Summer 2023

From Information Overload to Knowledge Graphs: An Automatic Information Process Model

Eve Huang Thullen
Claremont Graduate University

From Information Overload to Knowledge Graphs:
An Automatic Information Process Model

By
Eve Huang Thullen

Claremont Graduate University
2023

Approval of the Dissertation Committee

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Eve Huang Thullen as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Information System & Technology.

Wallace Chipidza, Chair
Claremont Graduate University
Assistant Professor of Information System & Technology

Samir Chatterjee, Professor
Claremont Graduate University
Fletcher Jones Chair of Technology Design & Management

Chinazunwa Uwaoma, Professor
Claremont Graduate University
Assistant Professor of Information System & Technology

Abstract

From Information Overload to Knowledge Graphs:
An Automated Information Process Model
By
Eve Huang Thullen

Claremont Graduate University: 2023

Continuously increasing text data such as news, articles, and scientific papers from the Internet have caused the information overload problem. Collecting valuable information as well as coding the information efficiently from enormous amounts of unstructured textual information becomes a big challenge in the information explosion age. Although many solutions and methods have been developed to reduce information overload, such as the deduction of duplicated information, the adoption of personal information management strategies, and so on, most of the existing methods only partially solve the problem. What's more, many existing solutions are out of date and not compatible with the rapid development of new modern technology techniques. Thus, an effective and efficient approach with new modern IT (Information Technology) techniques that can collect valuable information and extract high-quality information has become urgent and critical for many researchers in the information overload age.

Based on the principles of Design Science Theory, the paper presents a novel approach to tackle information overload issues. The proposed solution is an automated information process model that employs advanced IT techniques such as web scraping, natural language processing, and knowledge graphs. The model can automatically process the full cycle of information flow, from information Search to information Collection, Information Extraction, and Information Visualization, making it a comprehensive and intelligent information process tool. The paper presents the model capability to gather critical information and convert unstructured text data into a structured data model with greater efficiency and effectiveness. In addition, the paper presents multiple use cases to validate the feasibility and practicality of the model. Furthermore, the paper also performed both quantitative and qualitative evaluation processes to assess its effectiveness. The results indicate that the proposed model significantly reduces the information overload and is valuable for both academic and real-world research.

Dedication

I humbly dedicate this piece of work to my Lord and my loving family:

My Lord, the source of wisdom, thanks for revealing the true knowledge to me.

My Lord, thanks for your kindness and generosity. My two children Bjorn and Edwards, the precious gifts from you, are my spiritual support and constant motivations for this work

Acknowledgements

Words cannot express my gratitude to my Chair Committee Dr. Chipidza for his invaluable patience and feedback. I also could not have undertaken this journey without my defense committee members, Dr. Chatterjee, and Dr. Uwaoma who generously provided their knowledge and expertise.

Special thanks to my father and my husband, this research would not have been done without their endless patience and support. I would also like to thank my sister and her husband Dr. Yu, thanks for their editing help.

I am also grateful to my classmates and friends, for their editing help, late-night feedback sessions, and moral support. Special thanks to my friend Dr. Negritto and Dr. Al-kayed, their expert feedback is essential for my research.

Table of Contents

List of Figures

Chapter 1. Introduction

1.1 Information Explosion Age

The term Information Explosion refers to the unprecedented growth of digital information that has occurred in recent years, particularly with the rise of the Internet. A large amount of digital information has been created by the rise of the digital age, such as news, articles, bloggers, scientific news and so on. This exponentially increasing information not only grows simply by quantitative but also a qualitative way to create new knowledge and change the way we understand the world (IORG, 2022).

During the period 2010 to 2020, data interactions surged by an astounding 4,900% (Djuraskovic, 2021). It means that over the past eleven years, the overall volume of data generated, collected, replicated, and consumed worldwide has witnessed a nearly forty-eight-hundred-fold increase. This unprecedented expansion of data usage has led to a data escalation from 1.2 trillion gigabytes to 59 trillion gigabytes (Djuraskovic, 2021). There were 40 zettabytes of data created in 2020 alone, and the exponential growth of data shows no signs of slowing down - approximately 1.7 megabytes of data produced per second by each internet user worldwide (Djuraskovic, 2021). In 2021, the overall amount of data generated worldwide was estimated to be around 79 zettabytes. This volume of data is expected to double by 2025 according to forecasts by Statista (Taylor, 2022).

Information Accumulated as Growing Tech Stacks and Time Progressed

There are many reasons behind the information explosion. The primary cause of the information explosion is attributed to the burgeoning population of Internet users. The increasing number of individuals who can access the Internet has led to a significant rise in the production of -information - anyone with an internet connection can create and publish information in various formats, from news to articles, research papers and so on. The second cause of the information explosion is the accumulation of historical data on the Internet. The production of data on the Internet has resulted in a gradual accumulation of historical data, which has transformed the Internet into an extensive information repository. As time passes, the amount of historical data on the Internet continues to mount, further adding to the information explosion (Solis, 2020).

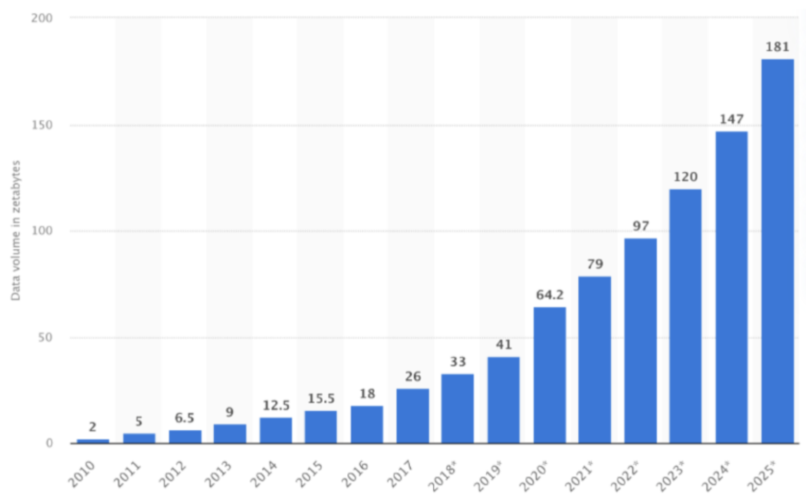


Figure 1. "Total Data Volume Worldwide 2010-2025." Statista, Sep 8. 2022

In summary, the proliferation of technology stacks, data silos, and the passing of time has resulted in a significant increase in data production. With so much information available, filtering and prioritizing essential information can be challenging and difficult.

1.2 Information Overload Challenges

The rapid and exponential growth of digital information has undoubtedly presented many opportunities to researchers. Nonetheless, it has also brought a host of challenges and problems (IORG, 2022). One of the primary challenges associated with the information explosion is information overload.

Bertram Gross, a professor at Hunter College, in his 1964 work The Managing of Organizations explained, "Information overload occurs when the amount of input to a system exceeds its processing capacity. Decision makers have limited cognitive processing ability. Consequently, when information overload occurs, it is likely that a reduction in decision quality will occur" (Solis, 2020, p.1). Information overload occurs "when the supply exceeds the capacity. Dysfunctional consequences ... and a diminished decision quality are the result" (Eppler & Mengis, 2004, p.32). Information overload is "therefore having more information than one can acquire, process, store, or retrieve"(Brennan, 2011, p.3).

The definition of information overload has undergone expansion over time, encompassing two added scenarios in contemporary usage (Huettich, 2020). These scenarios pertain to: (1) an insufficient amount of time for processing available information effectively or (2) an excessive and irrelevant abundance of information that surpasses the limits of our cognitive processing capabilities.

A notable issue in Internet-generated information is that a massive proportion of data produced on the Internet is unstructured text, making it difficult to process, compare and evaluate the information quickly and methodologically (Solis, 2020). Moreover, the lack of clear organizational structure and inadequate indicators of relationships among textual information worsens the challenge. "Post-modern society witnessed enormous changes in technology and innovations. As a result, quick access to relevant information became a nightmare. The digital media contents became abundant, and it became impossible to retrieve useful information when needed" (Renjith R., 2017, p.2).

A consequence of the information overload problem is "the paradoxical situation that, although there is an abundance of information available, it is often difficult to obtain useful, relevant information when it is needed" (Edmunds & Morris, 2020, p.1). And "this is arguably one of the biggest challenges for knowledge workers in this age of technology-driven explosions of information availability" (Brennan, 2011, p.3).

Overall, the digital era has brought about an exponentially increasing amount of available information. This has undoubtedly created vast opportunities for researchers and individuals, while it has also raised a challenge called information overload. The human brain has a finite capacity to process information, and this created two main challenges in the context of information overload. The first challenge is the information-gathering problem. With a vast amount of unstructured information available, locating useful data becomes difficult. Advanced skills are necessary to sift through the deluge of information to identify relevant and valuable data. Finding such data, therefore, becomes a significant challenge. The problem is compounded by the fact that the acquisition of the latest information is a never-ending process, and the sheer amount of information available is constantly increasing. The second challenge is the information-coding problem. Given the abundance of unstructured data available, it is challenging to absorb and store the information effectively. The coding of data is a resource-intensive and time-consuming process. Researchers encounter difficulties in managing large datasets of unstructured text information, which need a substantial amount of time for information extraction and analysis.

In conclusion, collecting valuable information from vast amounts of unstructured data is a significant challenge. The amount of data available can be overwhelming, and traditional information gathering and coding approaches may not be sufficient. Developing new and advanced methods for researchers to cope with the information explosion age is urgent.

1.3 Why does it matter?

The growth of the internet and the development of added information technology have contributed to the deluge of available data, which has made it increasingly challenging to gather, code, and analyze information effectively. Simultaneously, time management requirement is the most difficult and demanding element when the strategic intelligence process is considered (McDowell, 2008). Researchers face a tight timeline, making it difficult to deliver accurate results. To reduce the information overload and reduce anxiety and stress, as well as save time and labor, the design of an effective and efficient information process model to collect valuable data as well as extract high-quality information becomes urgent and critical in the information overload age.

Effective data collection is a crucial factor in reducing information overload. Gathering data correctly can streamline the information acquisition process, enabling researchers to identify high-quality information quickly. Researchers are often inundated with a large amount of data, and it can be challenging to distinguish between valuable and irrelevant information. Effective data collection strategies can significantly reduce this problem. Developing effective information coding techniques is equally critical in dealing with information overload. The coding process is time-consuming, and researchers may struggle to handle large datasets of unstructured text information, which impedes effective data analysis. Thus, it is essential to have reliable data coding methods to enable researchers to analyze and interpret data accurately.

In summary, information overload is a significant challenge arising from the increasing data availability in the digital age. To mitigate the information overload problem, it is critical to develop an effective information gathering and coding method or model. By reducing the burden of information overload, researchers can save time and effort and deliver accurate and precise results promptly.

1.4 Research Questions:

In the earlier session, we discussed that while the availability of digital information has brought significant benefits, it has also created challenges. Information overload is one such challenge that arises from the sheer volume of unstructured data available. It presents challenges in both the information gathering and coding processes. Overcoming these challenges requires developing advanced skills, innovative techniques, and novel approaches for information processing. To solve those challenges and problems, we proposed the following research questions:

1. In the information explosion era, how new technologies can help us to gathering textual information in a more efficient way?
2. In the information overload age, how new technologies can help us to coding unstructured text information to structured data in a more effective way?

Chapter 2. Literature Review

2.1 Prior Research

Maintaining Status Quo

The problem of information overload is not realistically solvable as it is always present. Instead, maintaining the status quo may be the more practical approach to dealing with this issue (Wilson, 1997). The issue of information overload is primarily due to the abundance of professional journals. According to Laskin, even if all the information contained in these journals was relevant, the sheer volume of material would still be overwhelming. Nevertheless, Laskin argued that the time spent reading these journals would not be considered wasted. However, he noted that reading duplicate information, encountering deficient data, and unsupported conclusions could result in wasted time. To combat this, Laskin suggested that users focus on concepts and principles rather than details and data (Laskin, 1994). Klassen, Jadad & Madher (1998) also agree that it is impossible to keep up to date with literature, especially medical research literature. They proposed a solution: systematic reviews and a scientific process to reduce information overload (Klassen et al., 1998). The systematic review is compiled by undertaking a comprehensive search for relevant studies, which are then appraised and synthesized according to a predetermined and explicit method. This provides the ability to replicate the research (Edmunds & Morris, 2020).

Strategies Focused

Hemp (2009) in the article Death by Information Overload stated that researchers need to change their mind-set, by seeking help from personal-productivity experts or by simply accepting that one could not respond to every distraction that flits across their screen. Similarly, organizations must change their cultures, for instance by setting up clear e-communication protocols (Hemp, 2009).

Brennan (2011, p.133) proposed a method called the scientific management of information overload and suggested three key principles to reduce information overload, the need for direct observation, the standardization of tasks, and the elimination of waste. Based on the analysis presented here, the following propositions are offered "to address the problems of information overload, the individual must first use direct observation to understand the scope of the problem. Cross-referencing an individual's roles and information actions provides a catalog of information needs, tracked over time. Applying gap analysis to the catalog of information needs will enable the individual to reduce the overload and improve the quality of information use. Standardization of information handling is needed to identify sources of waste in information processes. Reducing waste in information processing will further alleviate information overload" (Brennan, 2011). "If technology is the primary cause of information abundance, technology too has solutions. Intelligent use of technology will eliminate the actual problem for sure" (Renjith R., 2017, p.82). Renjith offered multiple strategies to alleviate information overload and get access useful information, "prioritize the information needed; identify the unwanted information; continuous observations; keep a disciplined and organized mind; planning before retrieving information; create a precise idea about what is going to access; use specific terms to search; write down the ideas immediately; learn search strategies; connect with known information; evaluate the retrieved content; Use only trusted sources " (Renjith R., 2017, p.83) and so on.

Using Intelligent Agent and Information Specialist

An information specialist is a trained professional who is responsible for managing information resources and providing support to individuals or organizations in accessing, analyzing, and utilizing information.It would seem an obvious solution to the problem of information overload in businesses to employ specialists in information handling to carry out the acquisition of relevant information-processing and packaging the information needed as appropriate (Edmunds & Morris, 2020).

Intelligent agents have been suggested to reduce the problem of information overload (Belfourd & Furner, 1997). Intelligent agents are smarter than average search tools for two principal reasons. Firstly, an intelligent agent acts with autonomy by making decisions on the basis of data it acquires about the environment, rather than as a result of direct instruction from the user; second, an intelligent agent has the facility to learn about individual personal preferences so that gradually it is able to predict the likelihood of items that will be of interest to the user.

Another solution is using an intelligence group, for example, there is a research group called Information Overload Research Group, which provides solutions and people or agents to help reduce the impact of Information overload for its members. It helps individuals or organizations to address the information overload issues and helps to facilitate conversations, collaboration, and networking among people. It also helps to educate organizations and individuals – about the economic and social costs of information overload, and cost-effective countermeasures; and spreading research-based solutions, which include best practices and technologies (IORG, 2022).

Technological Aids

Some technological aids could help to reduce information overload, such as e-mail management software, a message-volume regulation system for organizations, or even more-sophisticated software being developed by Microsoft, IBM, and others (Hemp, 2009). Also, software tools, such as filtering agents, automatic summarizers, or visualization algorithms, can help to process enormous amounts of information (Eppler & Mengis, 2004). Push-to-pull technologies by pushing notices of pre-selected information sources across the computer screen alerting users to new and updated information may be useful to reduce information overload (Edmunds & Morris, 2020). Some other techniques such as information filtering system or the recommendation system, also been used by reducing information overload, which relies on the historical interaction records of users' products and its own attribute information to explore the potential preferences and needs of users (Norton et al., 2015).

Limitations for traditional methods:

The issue of information overload has been a long-standing concern, many methods and solutions had been developed to reduce information overload, yet those solutions faced big limitations and challenges: The first major limitation is the generalizability of solutions. While many strategies have been developed to reduce information overload, they are often context-dependent and rely heavily on personal situations or context, making them difficult to generalize across different settings and populations. Consequently, there is a need for more flexible and adaptable solutions that can be applied across various domains. The second challenge is the rapid pace of technological advancement, which can render many existing solutions out of date and ineffective. Although some solutions such as information filtering systems or recommendation systems have been proposed, those systems are mostly implemented in the commercial industry for advertisement push to target users, plus, those systems are expensive. Therefore, it is essential to develop solutions that are adaptable and can keep pace with the ever-changing technological landscape. The third challenge pertains to the lack of details on how and what exactly IT techniques can do to reduce information overload. Despite the use of technological aids, such as data visualization tools or machine learning algorithms, they are vague and do not provide explicit details on how IT can reduce information overload or the specific IT tools that can be utilized to achieve this goal. Therefore, it is necessary to develop clearer and more detailed guidelines on the use of IT tools in reducing information overload.

Thus, the problem of information overload is a multifaceted issue that requires innovative and comprehensive solutions. The generalizability of solutions, the rapid pace of technological advancement, and the lack of details on how IT techniques can reduce information overload are some of the primary challenges that must be addressed to develop effective solutions. Consequently, researchers need to focus on creating flexible and adaptable solutions that can keep pace with technological advancements while providing clear and detailed guidelines on the use of IT tools in reducing information overload.

2.2 Theoretical Bases: Information Extraction and Knowledge Graph

Over the past few years, there have been significant advancements in artificial intelligence and machine learning technologies, leading to a noteworthy evolution in information extraction and information visualization techniques. Thus, information extraction and visualization has been identified as significant approaches to address the issue of information overload.

Information extraction (IE) refers to the process of automatically extracting structured information from unstructured or semi-structured data sources, such as text documents, websites, or audio and video recordings. “Information Extraction (IE) is a task of extracting pre-specified types of facts from written texts or speech transcripts, and converting them into structured representations (e.g., databases)” (Ji, 2009, p.1476). It involves identifying relevant information in a text and transforming it into a structured form that can be easily analyzed. Once information has been extracted, it can be organized into a knowledge graph.

“Knowledge graph is a knowledge base that uses a graph-structured data model or topology to integrate data. Knowledge graphs are often used to store interlinked descriptions of entities – objects, events, situations, or abstract concepts – while also encoding the semantics underlying the used terminology” (Ontotext, 2023, p.1). A Knowledge Graph (KG) is a powerful data structure that represents information in a graph format. It is composed of nodes (also referred to as entities) and edges (also called relationships), which represent the connections between these nodes. Additionally, KGs often include properties or attributes that provide further information about the nodes and edges. KGs are used to represent and organize information in a way that is easily navigable and understandable by machines and humans alike (Meyer, 2023). They can be used to represent several types of knowledge, including structured and unstructured data, and can be used in a variety of applications such as natural language processing, search engines, recommender systems, and more.

However, the process of information extraction and graph creation could be quite challenging. The traditional method of creating a graph, or more specifically, a Knowledge Graph, involves manual creation, which can be extremely time-consuming and requires a significant level of technical skills and domain knowledge. This often involves extracting information from various sources and organizing it into nodes and edges manually. This process can be error-prone and can lead to inaccuracies or inconsistencies in the resulting graph (Meyer, 2023).

Meyer (2023) presented an example about the Eiffel Tower and projected the description of Eiffel Tower from natural language to knowledge graph structure manually.

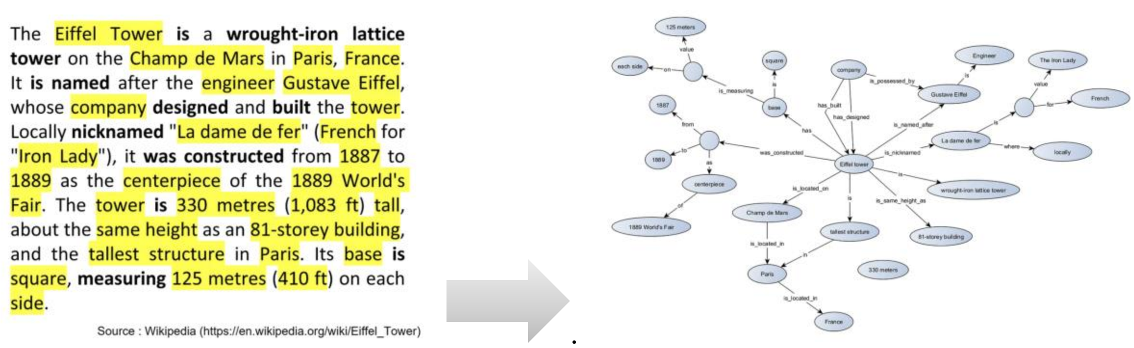
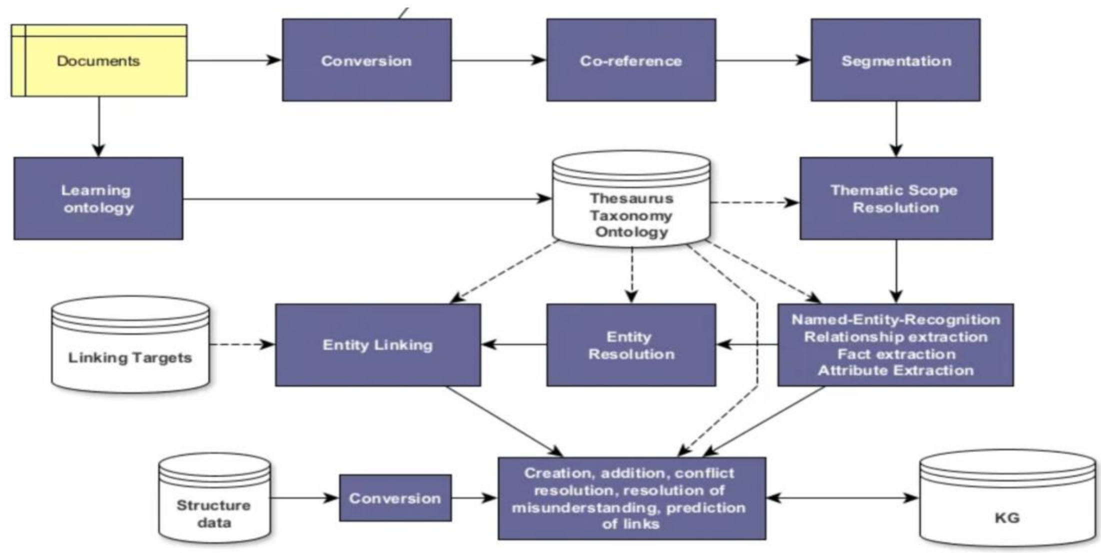


Figure 2. Textual Information to a Graph, Meyer 2023

Meyer then proposed a conceptual level of framework that could conduct information extraction and create a knowledge graph automatically (see fig. 2-2). Meyer's conceptual model described the key steps that transform documents to a knowledge graph, in which NLP algorithms are used in the model to automatically extract and organize information from various sources, such as text documents or databases and then the data is presented into a graph structure. The main goal of

Meyer's model is converting unstructured textual information to a structured knowledge graph data model automatically.



Sequence of automatic creation of a knowledge graph (author's diagram)

Figure 3. Automatic Creation of a Knowledge Graph, Meyer 2023

However, Meyer's model is limited to the conceptual level of the diagram only for transforming textual information to knowledge graph data structure. Meyer did not conduct further empirical research or apply the model to any use cases to examine the feasibility of the model but stated that "automatically creating a knowledge graph remains a complex and uncertain task, but it is technically achievable" (Meyer, 2023, p.5). To emphasize the impossible grail for this path, he elucidated several major challenges in the information extraction process that made it impossible to build automatic knowledge graphs. First, ambiguity is a prevalent issue. In textual information, one of the main challenges we encounter is the complexity of identifying a term within its context, both in its immediate linguistic surroundings and in a broader context. While most existing extraction models can label the type of entity detected, such as a person or place, they do not possess the capability to identify the same entity in different names. For example, my professor, Wallace Chipidiza, can be called Wallace, Dr. Chipidiza, or Professor Chipidiza. It can be very challenging for machines to recognize these entities are the same person. Second, the weight of words. As we engage with reading textual information, we bring to it our personal experiences and knowledge of the world. We apply our understanding of the field in question and approach the text critically. We assign different levels of importance to the information we encounter in the text. By assigning different weights to the words and relationships within the text, we are able to distinguish between crucial and insignificant information. Replicating this behavior in an automatic extraction program remains challenging. Additionally, the complexity of sentences. When sentences are written in a complex manner, they can be difficult to interpret and extract relevant information from. The use of complex grammatical structures, such as double negations, make the meaning of the sentence ambiguous or difficult to comprehend. It is essential to determine the maximum level of complexity of sentences that can be processed and integrated into a knowledge graph, which can be a challenging task. Also, handling complex sentences and defining appropriate strategies to process them can also be complicated. Therefore, accurately capturing the meaning of complex sentences remains a significant challenge in information extraction. Despite those challenges, Meyer stated that his diagram "is technically achievable."

2.3 Research Goals

Our main goal of this study continues to be transforming unstructured textual documents into a structured knowledge graph data model.

To do this, we will revamp Meyer's existing concept level of the automatic KG creation process model by emphasizing key constructs of the concepts while disregarding less significant ones. Simultaneously, we will introduce additional constructs to our updated or an improved model that makes the automatic information process to knowledge graphs possible. Most importantly, we will create and build actual IT artifacts of the model and will provide a detailed explanation of each IT element involved in the process, along with the methodology employed to execute each component. This approach will not only clarify the actions required and the reasons behind selecting these components, but it will also demonstrate the steps necessary to make it operational. What is more, we will conduct several real-world case studies to examine the feasibility of our model and demonstrate our model's practicality. Also, we will evaluate the model's performance.

Research Goal: from unstructured textual information to Knowledge Graphs Data Structure

	Meyer's Model	Improved Model
Conceptual Framework	YES	YES
Model Design: System IT Artifacts How, What, Why, Practicable	NO	YES
Model Demonstration: Use Cases, Real world Scenarios	NO	YES
Model Evaluations: Performance of the model	NO	YES

Figure 4. Meyer's Model vs. Projected Model

Chapter 3. Research Methodology

3.1 Design Science Research Theory

Research methodology can be defined as a systematic process that involves the use of data to seek answers to a specific question, solve a particular problem, or gain deeper insights into a phenomenon (Hevner & Chatterjee, 2010). Research methods can be divided into two primary types: descriptive research and prescriptive research (Hevner & Chatterjee, 2010). Descriptive research is aimed at gaining a better understanding of the nature of information technology (IT) and corresponds to knowledge-producing activities in natural science. On the other hand, prescriptive research is focused on enhancing IT performance and is categorized as a knowledge-using activity that corresponds to design science.

Science can be divided into two categories: natural science and design science (March & Smith, 1995). Natural science refers to a body of knowledge that encompasses a specific class of objects or phenomena in the natural or social world. The primary goal of natural science is to describe and explain how objects or phenomena behave and interact with each other. The objective of natural science is to develop specialized concepts or languages to explain phenomena, culminating in the formulation of theories, which serve as a deep principled explanation of phenomena. The key activities of natural science are discovering and justifying (March & Smith, 1995). Discovery is the process of generating or proposing scientific claims, while justification involves testing such claims for validity. However, design science is a discipline aimed at creating artifacts that serve human purposes. The fundamental goal of design science is problem-solving and providing valuable solutions to end-users, and the value of its products is assessed against criteria of usefulness and effectiveness. The center of design science is the concept of "design," which plays a vital role in fields such as architecture, engineering, and urban planning. Unlike natural science, the primary focus of design science is not on discovering general theoretical knowledge, but on producing and applying knowledge of specific tasks or situations to create effective artifacts (March & Smith, 1995). Design science comprises two basic activities: building and evaluation. Building involves constructing an artifact for a particular purpose, while evaluation is the process of determining how well the artifact performs (March & Smith, 1995).

In this study, we use Design Science Research (DSR) as our research methodology, in which we will create or build effective artifacts to help reduce the information overload problem. Based on Meyer's automatic knowledge graph creation process, we will build an improved automatic information extraction and KG process model. We will apply a real-world case study to our model and conduct an evaluation using real-world surveys to examine its performance.

3.2 Research Delivery and Research Process

3.2.1 Research Outputs

There are four possible outcomes for Design Science Research (DSR), which include constructs, models and frameworks, methods and algorithms, and instantiations (March & Smith, 1995). Constructs serve as the foundation for defining and communicating problems and solutions in a specific domain. They are abstract concepts that are intentionally selected or created to explain a particular phenomenon and form the vocabulary of the domain. For example, constructs may include notions of entities, objects, and data types that describe problems within the domain and specify their solutions. Models and frameworks use constructs to represent real-world situations, such as the design problem and its solution space (Simon, 1996). Models are useful in understanding problems and solutions, as they often show the relationship between the different components of a problem and its corresponding solution. They allow for the exploration of how design decisions and changes could impact the real world. For instance, constructs can be combined to represent problem and solution statements, and the focus of models is their usefulness. Examples of such models include use case scenarios and storyboards (March & Smith, 1995). Methods and algorithms define the processes used to develop and implement the artifact. They can range from formal mathematical algorithms that explicitly define the search process to informal, textual descriptions of best-practice approaches. An algorithm can be described as a set of steps used to perform a task or solve a problem, based on a set of underlying constructs and models of the solution space (March & Smith, 1995). Instantiations and demonstrations show that constructs, models, or methods can be implemented in a working system (Hevner et al., 2004). Instantiations and demonstrations provide tangible evidence that an artifact is feasible and can be assessed concretely in terms of its suitability for its intended purpose. Additionally, they demonstrate the realization of the artifact in its environment and confirm the feasibility and effectiveness of the models and methods it contains (Hevner et al., 2004).

In section 2.2, we explored Meyer's automatic knowledge graph mode, and our goal of this study is to redesign Meyer's model and add new IT artifacts to the model. Of the four outputs of DSR, our delivery will belong to Models and Frameworks.

3.2.2 Research Process

This paper's structure will follow the Design Theory Six Step Process (Peffers et al., 2007), which consists of six steps guiding the development of an innovative artifact to solve a specific research problem.

First, problem identification and motivation, which involve defining the specific research problem and justifying the value of a solution. The goal is to clearly identify the problem and explain why it is important to solve it. We had described the information overload problem in the previous introduction chapter. Second, define the objectives for a solution, which are inferred from the problem definition and knowledge of what is possible and feasible. The goal is to articulate a set of clear and measurable objectives that the artifact should achieve. We discussed in the previous chapter that the study's main goal is to convert unstructured textual documents into a structured knowledge graph data model. To accomplish this, we will modify Meyer's existing automatic KG creation process model by emphasizing essential constructs of the concepts while disregarding less significant ones. Meanwhile, we will bring in new constructs to our improved model that made the automatic knowledge graph creation possible. Third, design and development involved in creating the artifact. The goal is to design and develop an innovative solution that meets the identified objectives. Our approach includes creating and building actual IT artifacts of the model, along with a detailed explanation of each IT element involved in the process and the methodology employed to execute each component, which will be explored in the next chapter. Next is demonstrations and case study for the model. In this step, the use of the artifact to solve one or more instances of the problem is demonstrated. The goal is to display the feasibility of the artifact in solving real-world problems. We will conduct multiple real-world case studies to examine the practicality and feasibility of our model in chapter 5. Then, evaluation, which involves observing and measuring how well the artifact supports a solution to the problem. The goal is to evaluate the effectiveness of the artifact in achieving the identified objectives. The evaluation process will be explained in chapter 6. The last step is to communicate the results of the research through publication.

By following the Design Theory Six Step Process, this paper aims to develop an improved model with innovative artifacts that reduces information overload problems and contributes new knowledge to the field of study.

3.3 Conceptual Framework and Theoretical Bases

In the previous section of this paper, we provided an overview of the Design Science Research method. This discipline is dedicated to creating artifacts that serve human purposes, with a primary focus on problem-solving and providing value to end-users through effective design. The fundamental goal of Design Science is to create or build artifacts that can address specific problems. As discussed in this paper's introduction section, information overload is a significant problem that can negatively impact productivity. In this section, we will apply the Design Science Research method to rebuild Meyer's conceptual level automatic knowledge graph model that can mitigate information overload problems.

3.3.1 Conceptual Level Framework

In the second section of this paper, we discussed Meyer's research on the automatic creation of knowledge graph diagrams. Meyer's framework offers a valuable approach to generating knowledge graphs automatically by utilizing the Information Extraction process. However, the diagram is overly complex. To improve the practicality of Meyer's approach, we believe with some modifications, we can make it more practical and better suited to our specific needs. Thus, we have decided to simplify Meyer's framework and integrate some new IT artifacts into our own framework.

Thus, our new model will incorporate Meyer's information extraction process, which we have refined by eliminating excessive detail levels such as Co-reference, Segmentation, Thematic Scope Resolution, Thesaurus Taxonomy Ontology, etc. Instead, we will consider Information Extraction as a unified component of our new model. In terms of the knowledge graph aspect, information visualization will be the primary component, with the knowledge graph serving as the fundamental instantiation of the information visualization component.

To address our research questions regarding how we can use modern technologies to design more efficient information gathering methods for users in the information explosion era, we will add two more components to our model artifact: information search and information collection.

Therefore, our proposed new model will encompass the entire spectrum of information management, from the Information Search skills to Information Collection techniques, to methods of Information Extraction, and to Data Visualization of insights. In this way, we expect to build an information streamline that flows through the model smoothly.

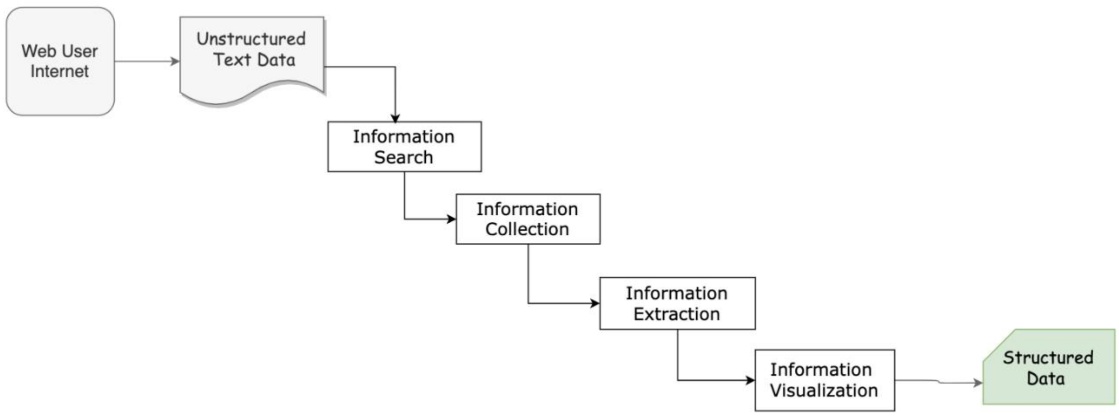


Figure 5. Conceptual level Framework: An Automatic Information Data Flow

It is important to keep in mind that although we reconstructed Meyer's model for creating automatic knowledge graphs, we still face challenges that cannot be ignored. When it comes to textual data, it cannot be directly fed into machine learning algorithms, as these algorithms only comprehend numerical data. Therefore, there is a need to convert textual data into a numerical format before it can be processed by these algorithms (Shankar, 2022). Thus, one of the most challenging issues we will face in our research is the Information Extraction process. "Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents and other electronically represented sources. In most cases this activity concerns processing human language texts by means of natural language processing (NLP)" (Wikipedia, 2023).

Therefore, to ensure that we make progress and develop our IT architecture effectively, we will begin with introducing some theoretical knowledge bases first.

3.3.2 Theoretical Bases for IT Artifact:

Our objective is to construct a highly efficient model that incorporates modern IT methodologies as its components. We have carefully chosen several crucial IT techniques to incorporate into our model, with the aim of improving its efficiency and effectiveness. Chapter 4 will provide further elaboration on our rationale for selecting these IT components as the artifacts of our model. Also, further explanation of their function and implementation methodology will be explored.

Information Search – Open-Source Intelligence (OSINT)

Open-source intelligence (OSINT) is a rapidly evolving field of intelligence that has gained prominence in recent years. In the intelligence community, the term "open" refers to overt, publicly available sources. "Open-source intelligence (OSINT) is an intelligence that is produced from publicly available information and is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement" (109th CONGRESS, 2006). Open-source intelligence is a "multi-methods methodology for collecting, analyzing, and making decisions about data accessible in publicly available sources to be used in an intelligence context. In the intelligence community, the term open refers to overt, publicly available sources" (Wikipedia, 2022). OSINT is defined as an information search derived from publicly available information and collected, analyzed, and distributed promptly to a relevant audience to fulfill a specific intelligence requirement. It involves collecting information from a wide range of publicly available sources such as social media platforms, news outlets, academic journals, and government websites, etc. In the forthcoming section on system artifacts, we will offer more in-depth insights into OSINT, including its key components, methodology, tools, and framework.

Information Collection - Web Scraping

Web scraping involves bots to extract data and content from websites. It accesses the underlying HTML code and the data stored in a database. It is a straightforward method used to gather vast amounts of data from websites automatically (Agrawal, 2021). The data can be categorized into three types: structured, unstructured, and semi-structured. Websites tend to store all these types of data in an unstructured format, and web scraping is a technique that can be employed to collect this unstructured data from websites and store it in a structured manner. In the next section on system artifacts, we will delve into further details regarding the methods and tools used in Web Scraping.

Information Extraction – Natural Language Process NER and RE

Information Extraction is a process that involves the parsing of unstructured data and extracting essential information to transform it into more structured and editable data formats (Kurama, 2021). In natural language processing, an Information Extraction (IE) system is utilized to take natural language text as input and produce structured information by certain criteria. There are several subtasks of Information Extraction, including Named Entity Recognition, Named Entity Linking, Relation Extraction, and Knowledge Base Reasoning (Singh, 2018). These subtasks serve as the foundational building blocks for various high-end NLP tasks, such as Machine Translation, Question-Answering Systems, Natural Language Understanding, Text Summarization, and Digital Assistants, etc. Consequently, we will employ Information Extraction techniques such as Named Entity Recognition, Named Entity Linking, and Relation Extraction in our artifacts. Further details on the utilization of these techniques will be provided in the subsequent chapters.

Information Visualization and Storage – Knowledge Graph Theory

At section 2.2, we had discussed the concept of Knowledge Graph Theory. Knowledge Graphs are a form of data management that incorporates features from databases, graphs, and knowledge bases.

They are composed of entities and their connections, which are represented as nodes and edges (Borhade, 2021)). One of the key advantages of Knowledge Graphs is their ability to model one-to-many relationships (Borhade, 2021). Traditional Legacy or relational data storage faces several significant challenges that can impede data management. First, relational databases suffer from rigid schemas, making it difficult to adapt to changes in complex enterprise environments, which can become messy and difficult to organize. Also, data stored in legacy systems can be challenging to understand and manage, requiring complex join queries, and indexing data can become expensive. Additionally, data stored in legacy systems may lack business knowledge and relationship information, making it challenging to extract valuable insights. Finally, legacy data stores can take up to 5 times more disk space than modern graph databases, and relationships and joins need to be recalculated for each query, further complicating the problem (Borhade, 2021). To overcome these challenges, Knowledge Graph is a powerful and innovative data structure model for future knowledge base storage.

Conclusively, this research utilizes the output from the Design Science method: models and frameworks. By leveraging the output, we propose an enhanced model and construct an IT artifact that can effectively mitigate the impact of information overload. In accordance with Peffers' (2007) Six Step Process of Design Theory, the problem of information overload and project output were already defined in the preceding section. The subsequent portion of this paper will explain more detailed information about the model artifacts and each IT component. Building upon current research, we aim to deliver a comprehensive and innovative solution that leverages modern IT techniques to address the problem of information overload.

Chapter 4. Model Design: System IT Artifacts

In the previous chapter, we presented the fundamental IT components that will be used as the artifact in our model, such as Open-Source Intelligence (OSINT), web scraping, natural language processing (NLP), and knowledge graph. In this chapter we will discuss more detailed level information about what the functions of those IT components are and how those IT components can serve our research purpose, as well as the implementation methodology for each component.

The first component of our IT artifact is to use OSINT to conduct information search. In this chapter, we will present a comprehensive overview of OSINT, covering its definition, historical background, framework, processes, and related techniques. Additionally, we will discuss various tools and methods that can be employed to enhance the effectiveness of OSINT. The second component of our IT artifact is web scraping, which is a widely employed technique for data collection. This chapter we will offer a comprehensive explanation of web scraping and its fundamental principles and guidance. Additionally, we will delve into the web scraping process, highlighting its key elements and considerations. Furthermore, we will provide an overview of some of the most popular web scraping tools available today, which can be used to optimize the effectiveness of the web scraping process. The third component of our model artifact is natural language processing (NLP) for information extraction process, in this chapter, we will introduce the Natural Language Processing (NLP) component of our IT artifact, which comprises Name Entity Recognition and Relation Extraction. We will provide a comprehensive explanation of NLP and its various applications in the IT field. Additionally, we will explore the different tools and techniques that can be utilized to implement NLP. Last, we will explain knowledge graph components, which is a vital component of our IT artifact and is primarily used for information visualization and storage. We will discuss available methods and tools that can be employed to leverage Knowledge Graph for visualizing complex data structures and relationships.

4.1 Artifact: Information Search, Open-Source Intelligence (OSINT)

Information search, also known as information seeking or querying, refers to the process of seeking specific information in response to a well-defined and articulated need. Information searching involves targeted and focused efforts to locate relevant information through various channels such as search engines, databases, or libraries. Information search can be aided by search techniques such as Boolean logic, keyword searching, and advanced search operators. The goal of information searching is to find the most accurate and reliable information to meet your information needs (Fulton & McGuinness, 2016). There are many information search skills and strategies, one of the most valuable skills and strategies for searching information is OSINT. OSINT techniques can include conducting advanced searches on search engines, using specialized tools and software to gather and analyze data, monitoring social media and online forums, and analyzing public records and financial information. OSINT is particularly useful in today's digital age where there is an abundance of information available online (Minner, 2018). The use of OSINT can be traced back to the Cold War era when the US (United States) intelligence community began using open-source information to support military and intelligence operations (Dunnigan, 2002). OSINT was first introduced as an intelligence tool during World War II, when the United States established the Foreign Broadcast Information Service (FBIS) to monitor publicly available information that supported its troop operations at that time (Bielska, 2020). The FBIS initially relied on traditional print media such as newspapers, magazines, and radio broadcasts to gather information. In the 1980s, OSINT was reintroduced with the advent of the web, social media, and digital services. The Director of National Intelligence (DNI) announced the creation of the Open-Source Center OSC (Open-Source Center), which was tasked with collecting information available from both online and offline public sources, which was previously done by the FBIS. And then, the Intelligence Reform and Terrorism Prevention Act proposed to reform the intelligence activities of the US government. This Act merged the FBIS and other related research entities into one body. This organization is now called the Open-Source Enterprise and is managed by the CIA. The merger of the FBIS and related research entities into the Open-Source Enterprise managed by the CIA became modern OSINT (Bielska, 2020).

4.1.1 OSINT Methods and Tools

OSINT methods and tools refer to a diverse range of publicly accessible information sources (Hassan & Hijazi, 2018). These include the internet, which encompasses a vast array of platforms such as forums, blogs, social networking sites, video-sharing sites like YouTube.com, wikis, Whois records of registered domain names, metadata and digital files, dark web resources, geolocation data, IP addresses, people search engines, and anything that can be found online. In addition to the Internet, traditional mass media such as television, radio, newspapers, books, and magazines are also considered as sources of OSINT. Specialized journals, academic publications, dissertations, conference proceedings, company profiles, annual reports, company news, employee profiles, and photos and videos (including metadata) are further examples of OSINT sources. Geospatial information, including maps and commercial imagery products, is also included as a key component of OSINT.

Necessary skills of OSINT, such as using advanced search engine queries; individuals learn how to identify essential information quickly and fast (Hassan & Hijazi, 2018). Additionally, some skills may refer to how to locate information buried in the deep web, access and navigate the Dark Web, and collect intelligence from the Dark Web, and so on.

4.1.2 OSINT Framework

To make OSINT more efficient and effective, there is an OSINT framework that focuses on gathering information from free tools and resources. The framework aims to provide a comprehensive guide for OSINT practitioners to find valuable information from various sources without incurring considerable time (JNordine, 2022). The framework provides guidance on how to use different tools or platforms to gather information, including advanced search techniques and tools such as Social Mention and Hootsuite. More details are found at the following link: https://osintframework.com/

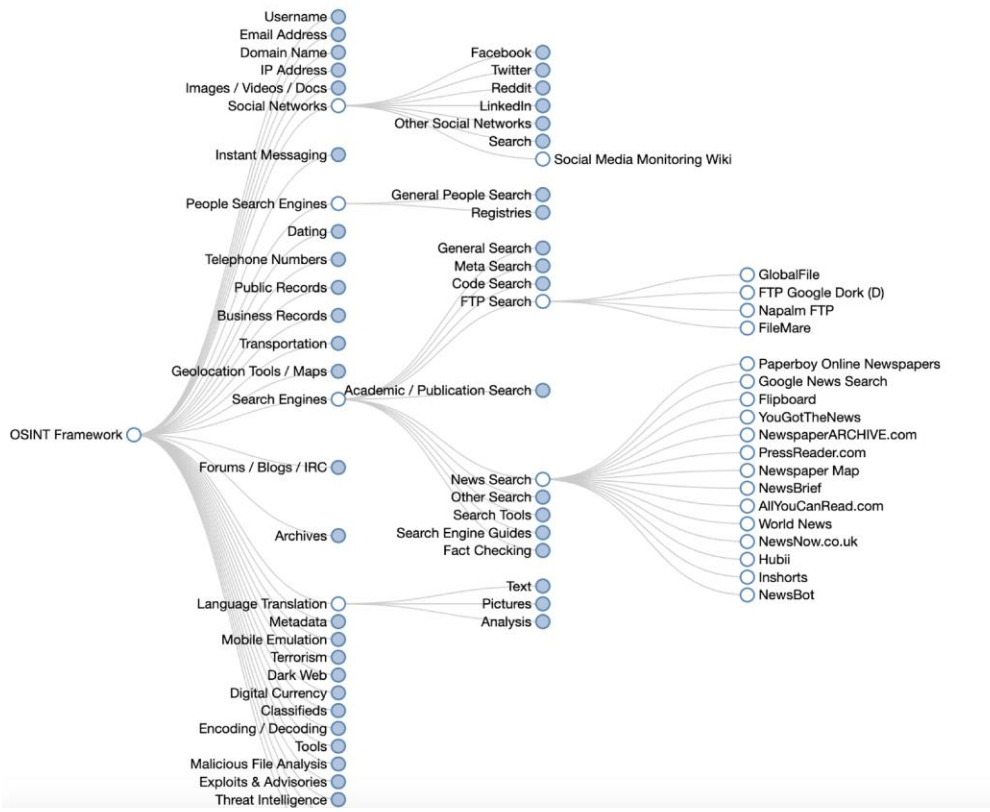


Figure 6. OSINT Framework, developed by Justin Nordine

Overall, OSINT is a particularly useful strategy for information search as well as a valuable repository that provides various methods, skills, and tools for information search.

4.2 Artifact: Information Collection, Web Scraping

In section 3.3, we described web scraping, also known as data scraping, a technique used to automatically extract enormous amounts of data from websites. Web scraping is a straightforward method for automatically collecting enormous amounts of data from websites (Agrawal, 2021).

Web scraping has appeared as a widely used and potent tool for extracting data from websites. Its ability to gather information swiftly and efficiently, regardless of the data's size, makes it a favored choice among researchers and data analysts. The various advantages of web scraping make it an essential aspect of the information collection process. One of the primary benefits of web scraping is its ability to extract data that is impossible to collect manually (Patel, 2022). Web Scraping offers a noteworthy advantage of saving time and effort in data collection. Manual data extraction is a time-consuming task that can take days or even weeks, depending on the amount of data. Web scraping automates the data extraction process and with the help of a web scraping tool, users can extract any kind of data, including text, images, and multimedia files from any website and save it in a format such as CSV. This eliminates the need for manual data entry and copying, making data extraction simpler and more efficient and frees up valuable time and resources (Patel, 2022). Also, the extracted data can be saved in a format like csv that is easy to retrieve, analyze, and use. By providing the data in a CSV format, web-scraping tools simplify the process of data analysis, making it more accessible and user-friendly.

One of the most significant challenges of web scraping is the variety of website formatting and templates. Each website is different, and therefore, web scrapers need to inspect through website HTML to extract relevant information. This process can be time-consuming and requires many resources. The differences in formatting and templates can also result in inconsistencies in the extracted data, making it difficult to derive meaningful insights. The durability of web scraping is another challenge. Websites are constantly updated with new postings and formatting changes, making it challenging for web scrapers to extract accurate and relevant data continually. Once a web scraper is built and runs flawlessly, it does not necessarily mean that it will always run correctly. Developers need to continuously update and modify the scraper to ensure its effectiveness.

4.2.1 Web Scraping Method and Tools

The Web Scraping technique involves fetching and extracting data from a website's HTML code and transforming it into a structured format that can be easily analyzed. A typical web scraping process was described as having three phases. The first phase involves fetching the data via the HTTP protocol, which is an Internet protocol used to send requests and receive responses from a web server. The second phase involves the extraction of the data from the fetched HTML document. Finally, in the third phase, the extracted data is transformed into structured datasets (Persson, 2019).

There are several programming languages and tools available for building web scrapers, including Java, Python, Ruby, and JavaScript (Persson, 2019). However, Python has emerged as one of the most popular languages for web scraping due to its ease of use and availability of powerful libraries such as Beautiful Soup and Scrapy. Beautiful Soup is a Python library that is widely used for web scraping, and it allows for the collection of data from websites by accessing the underlying HTML code, in which developers can parse HTML and XML documents and extract useful information from them. Beautiful Soup also allows users to search for specific HTML tags or attributes and extract the data contained within them. It also provides advanced features such as navigating parse trees, handling different encodings, and handling poorly formed HTML, etc. Scrapy is another popular Python library for web scraping. It is a high-level web crawling and web-scraping framework that allows developers to write spiders to automatically extract data from websites. Scrapy is built on top of the Twisted asynchronous networking library, which allows it to operate at high speeds and handle large volumes of data. More details on Web Scraping tools can be found in Appendix: Web Scraping Tools.

4.2.2 Python BeautifulSoup

Given the popularity and versatility of Python libraries, our paper will focus primarily on Python tools and libraries for web scraping. BeautifulSoup is the most popular Python library used for web scraping and data extraction. It provides a powerful and flexible way to parse HTML and XML documents, making it an essential tool for many data-driven projects. There are several advantages to using BeautifulSoup, which we will discuss in detail below (Zenrows, 2023).

One of the main advantages of BeautifulSoup is its speed. It is designed to be fast and efficient, which makes it ideal for working with large datasets. When compared to other popular parsing libraries, BeautifulSoup often outperforms them in terms of speed and memory usage. This can make a significant difference when working with large datasets, allowing you to extract the data you need quickly and easily. Another advantage of BeautifulSoup is that it is beginner-friendly and easier to set up. The library has a simple and intuitive API that makes it easy to get started with web scraping, even if you have little or no experience with Python. The documentation is also professionally written and provides clear examples of how to use the library, which can help you get up and run quickly. In addition to being fast and easy to use, BeautifulSoup requires less time to run. This is because it does not load the entire webpage, but instead only loads the parts that you need. This can help to speed up your web scraping tasks and reduce the amount of time you spend waiting for your code to run. Further, BeautifulSoup can parse both HTML and XML documents, making it a versatile tool for working with a wide range of web content. This means that you can use it to extract data from web pages, RSS feeds, and other sources that use these formats. This flexibility makes it a valuable tool for many different applications. Finally, BeautifulSoup is easier to debug than other web scraping tools. This is because it provides clear error messages that can help you quickly identify and fix any issues with your code. This can save you a lot of time and frustration when working on complex web scraping projects.

In short, BeautifulSoup is a powerful and flexible library that provides many advantages for web scraping and data extraction. Its speed, ease of use, and ability to parse both HTML and XML documents make it an essential tool for many data-driven projects. Whether you are a beginner or an experienced programmer, BeautifulSoup is an excellent choice for your web scraping and data collection needs.

Thus, we will use the Python library Beautiful Soup as an example for our use case study. Here is an example to display how Python library Beautiful Soup work: A researcher would like to collect bio information of Elon Musk from the following website:

https://www.biography.com/business-figure/elon-musk

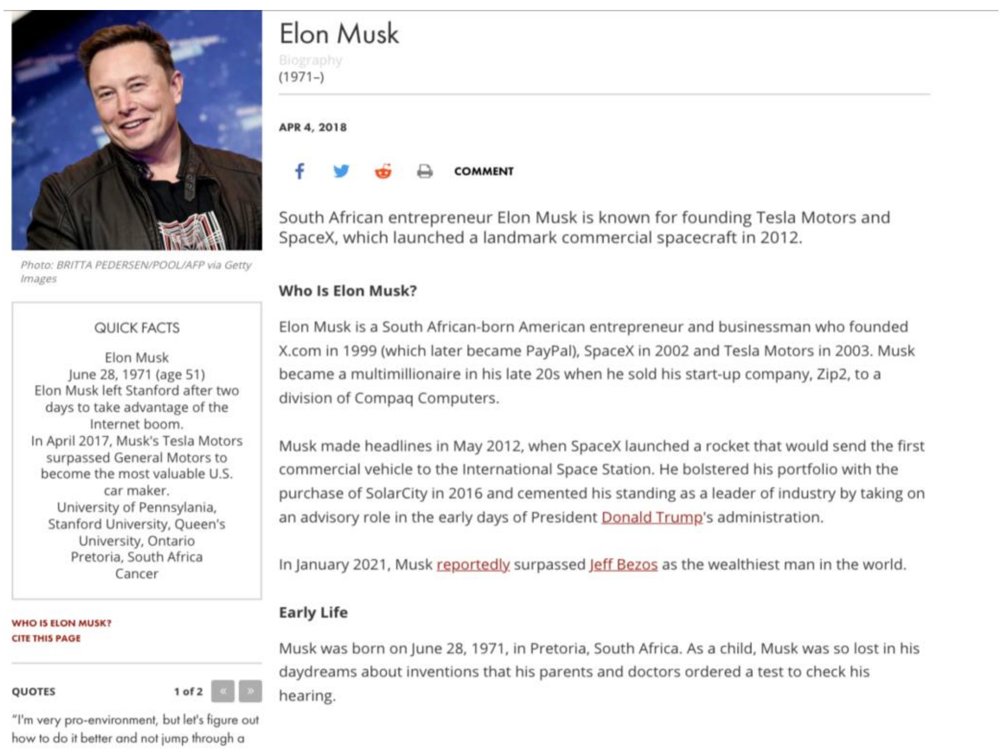


Figure 7. Elon Musk Biography

Through Python library Beautiful Soup, our scrapping results as below:

```
[[ Elon Musk is a South African-born American entrepreneur and businessman who founded X.com in 1999 (which later became PayP
al), SpaceX in 2002 and Tesla Motors in 2003.],
 [Musk became a multimillionaire in his late 20s when he sold his start-up company, Zip2, to a division of Compaq Computer
s.],
 [  Musk made headlines in May 2012, when SpaceX launched a rocket that would send the first commercial vehicle to the Intern
ational Space Station.],
 [He bolstered his portfolio with the purchase of SolarCity in 2016 and cemented his standing as a leader of industry by taki
ng on an advisory role in the early days of President Donald Trump's administration.],
 [In January 2021, Musk reportedly surpassed Jeff Bezos as the wealthiest man in the world.],
 [Musk was born on June 28, 1971, in Pretoria, South Africa.],
 [As a child, Musk was so lost in his daydreams about inventions that his parents and doctors ordered a test to check his hea
ring.],
 [At about the time of his parents' divorce, when he was 10, Musk developed an interest in computers.],
 [He taught himself how to program, and when he was 12 he sold his first software: a game he created called Blastar.],
 [In grade school, Musk was short, introverted and bookish.],
 [He was bullied until he was 15 and went through a growth spurt and learned how to defend himself with karate and wrestlin
g.],
 [Musk's mother, Maye Musk, is a Canadian model and the oldest woman to star in a Covergirl campaign.],
 [When Musk was growing up, she worked five jobs at one point to support her family.],
 [Musk's father, Errol Musk, is a wealthy South African engineer.],
 [ Musk spent his early childhood with his brother Kimbal and sister Tosca in South Africa.],
 [His parents divorced when he was 10.],
 [At age 17, in 1989, Musk moved to Canada to attend Queen's University and avoid mandatory service in the South African mili
tary.],
 [Musk obtained his Canadian citizenship that year, in part because he felt it would be easier to obtain American citizenship
via that path.],
 [In 1992, Musk left Canada to study business and physics at the University of Pennsylvania.],
 [He graduated with an undergraduate degree in economics and stayed for a second bachelor's degree in physics.],
 [After leaving Penn, Musk headed to Stanford University in California to pursue a PhD in energy physics.],
 [However, his move was timed perfectly with the Internet boom, and he dropped out of Stanford after just two days to become
a part of it, launching his first company, Zip2 Corporation in 1995.],
 [Musk became a U.S. citizen in 2002.],
```

Figure 8. Elon Musk Biography Python Web Scraping Results

Saving the scrapping textual paragraphs to .csv format:

0	Elon Musk is a South African-born American entrepreneur and businessman who founded X.com in 1999 (which later became PayPal), SpaceX in 2002 and Tesla Motors in 2003.
1	Musk became a multimillionaire in his late 20s when he sold his start-up company, Zip2, to a division of Compaq Computers.
2	Musk made headlines in May 2012, when SpaceX launched a rocket that would send the first commercial vehicle to the International Space Station.
3	He bolstered his portfolio with the purchase of SolarCity in 2016 and cemented his standing as a leader of industry by taking on an advisory role in the early days of President Donald Trump's administration.
4	In January 2021, Musk reportedly surpassed Jeff Bezos as the wealthiest man in the world.
...	...
166	Larry Page is an internet entrepreneur and computer scientist who teamed up with grad school buddy Sergey Brin to launch the search engine Google in 1998.
167	Entrepreneur and investor Paul Allen was best known for being one of the co-founders of Microsoft with Bill Gates.
168	Oscar Pistorius is a South African sprint runner who made history in 2012 as the first amputee to compete in track events at the Olympics.
169	He was later found guilty of murdering his girlfriend on Valentine's Day 2013.
170	Mark Zuckerberg is co-founder and CEO of the social-networking website Facebook, as well as one of the world's youngest billionaires.

Figure 9. Elon Musk Biography Web Scraping Results to CSV format

4.2.3 Some Challenges

Web scraping is an efficient and effective method for researchers to gather information, however, while conducting our own research, we encountered several other challenges that required additional attention and effort to overcome.

The quality of the extracted data is a significant challenge in web scraping - unreliable information and ungrammatical language of web scraping (Persson, 2019). Even if the information exists and is scrapable, it might not be correct. Grammatical errors and spelling mistakes can affect the parsing phase, leading to missed or falsely gathered information. The lack of quality control in web scraping can result in the extraction of unreliable information, leading to incorrect conclusions. Legal issues are also a significant challenge in web scraping. A research project that involves collecting enormous amounts of data via web scraping might accidentally compromise the privacy of a person. Researchers could match the data collected with another source of data, revealing the identity of the person who created the data. Web scraping can also raise copyright concerns, as it involves the extraction of copyrighted material from websites without permission.

Web scraping faces several challenges as we discussed above, including variety, durability, unreliable information and language, and legal issues. When implementing the model, researchers need to address these challenges to ensure that web scraping remains a reliable and effective means of data collection. Quality control measures, such as grammar and spelling checks, should be implemented to ensure the accuracy of the extracted data. Legal concerns, such as privacy and copyright issues, need to be considered when using web scraping in research and other applications. Overall, with careful consideration and attention to these challenges, web scraping can continue to be a valuable tool for data collection in our model.

4.3 Artifact: Information Extraction, NLP

In Section 3.3, a brief introduction was provided on the topic of Information Extraction. Information extraction is the process of automatically extracting structured information from unstructured and semi-structured machine-readable documents and other electronically represented sources (Wikipedia, 2023). This process typically involves natural language processing (NLP) to process human language texts. Information extraction is a critical component of various high-end NLP tasks, such as machine translation, question-answering systems, natural language understanding, text summarization, and digital assistants. The process of information extraction involves parsing unstructured data and extracting essential information to transform it into more structured and editable data formats (Kurama, 2021).

However, Information Extraction is an extremely challenging process. As mentioned in the given paragraph Conceptual Level Framework (see section 3.3), machine-learning algorithms only understand numerical data and are unable to comprehend text data. To address this challenge, researchers and practitioners in IE have been developing and using various natural language processing (NLP) techniques. These techniques involve the use of algorithms and models to process and analyze text data, such as syntactic and semantic analysis, named entity recognition, and sentiment analysis. NLP techniques have been instrumental in enabling IE algorithms to process and extract information from textual data effectively. The main challenges that hinder the extraction of useful information from unstructured data are its scalability, dimensionality, and heterogeneity (Adnan & Akbar, 2019). Also, the transformation of unstructured data into a structured format to improve its representation poses a significant question that must be addressed. In the following session, a more detailed exploration will be explored regarding information extraction methods, processes, and tools, as well as the many subtasks involved in the extraction process, including named entity recognition, relation extraction, and so on.

4.3.1 Named Entity Recognition (NER)

In Natural Language Processing, entity recognition and relationship extraction are the most important tasks. The term "entity" refers to a person, organization, event, location, or time that carries valuable information, which can be extracted from unstructured text. Named Entity Recognition (NER), also known as entity identification or entity extraction, is a natural language processing technique that automatically identifies named entities in a text and categorizes them into predefined categories. The categories of named entities include names of people, organizations, locations, times, quantities, monetary values, percentages, and more (*Named Entity Recognition*, 2020). One of the significant advantages of NER is its ability to perform information extraction from unstructured data. It enables the extraction of significant information to comprehend the subject of a text, and it can be used to collect essential information for storage in a database (*Named Entity Recognition*, 2020). Additionally, NER can help in discovering relationships between entities. By identifying and categorizing named entities, one can analyze the relationships between them, such as the relationship between a person and an organization. This relationship can help in understanding the structure of the text data and revealing insights that were hidden (ExpressAnalytics, 2022).

To facilitate the implementation of NER, there are several NER tools available for researchers. Stanford Named Entity Recognizer (SNER), Natural Language Toolkit (NLTK), and SpaCy are some of the commonly used NER tools. The SNER is a Java tool developed by Stanford University and is considered the standard library for entity extraction. It uses Conditional Random Fields (CRF) as a statistical model to identify and categorize named entities in text. SNER provides pre-trained models for extracting person, organization, location, and other entities, making it easy to get started with NER (*The Stanford Natural Language Processing Group*, 2023). The NLTK is a widely used tool for various NLP tasks, including named entity recognition. NLTK has its own classifier called ne_chunk, which can recognize named entities, and it also provides a wrapper to use the Stanford NER tagger in Python. More importantly, NLTK has an extensive community of users and developers who continuously improve the quality and availability of its tools (*Nltk :: Natural Language Toolkit*, 2023). SpaCy is a Python-based framework that is widely used by NLP practitioners due to its user-friendly interface and the availability of pre-trained models for NER tasks. It has an excellent statistical system that can be used to build customized NER extractors. Additionally, SpaCy can perform other NLP tasks such as dependency parsing and part-of-speech tagging (Spacy Universe, 2023).

For this study, we will implement SpaCy as our primary package.

Python SpaCy

"SpaCy is a free, open-source Python library that provides advanced capabilities to conduct natural language processing (NLP) on large volumes of text at high speed" (Domino Data Science Dictionary, 2023, p.1). The SpaCy library, launched in 2015 by Matthew Honnibal and Ines Montani, has become a popular tool in the NLP community due to its user-friendly interface and high processing speed. With a growing set of plug-ins and integrations, spaCy provides features for a wide range of natural language tasks. Its features range from tokenization, named entity recognition, dependency parsing, to sentence segmentation, and more (Singh, 2023). These features make spaCy an excellent choice for researchers and practitioners who need to process large volumes of text for various NLP tasks.

"SpaCy is designed specifically for production use and helps you build applications that process and "understand" large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning" (Spacy Universe, 2023b). SpaCy boasts an NER system that is characterized by its impressive speed and reliance on statistical analysis. This system is designed to assign labels to contiguous spans of tokens, which enables the identification and classification of entities within text data (Majumder, 2021). The default model or pertained model provided by spaCy is particularly noteworthy, as it can accurately identify a diverse range of named and numeric entities such as companies, locations, organizations, and products (Jain, 2018). Additionally, spaCy provides users with the ability to add arbitrary classes to the entity recognition system, which enhances the customization and specificity of the model. Users can also update the model with new examples, which contributes to its continued improvement over time.

4.3.2 Coreference Resolution

NER is the crucial task in natural language processing that involves identifying entities such as people, organizations, and locations in each text. However, in the process of NER, we often face the challenge of coreference resolution, which refers to the identification of all expressions in a text that refer to the same entity. Coreference resolution is the task of "finding all expressions that refer to the same entity in a text. It is a crucial step for a lot of higher-level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction" (*The Stanford Natural Language Processing Group*, 2023).

The challenge of coreference resolution in NER arises from the ambiguity of natural language. In many cases, multiple expressions in a text can refer to the same entity, and these expressions can take various forms, such as pronouns, proper nouns, and common nouns. For example, consider the following sentence:

"After leaving Penn, Musk headed to Stanford University in California to pursue a PhD in energy physics. However, his move was timed perfectly with the Internet boom, and he dropped out of Stanford after just two days to become a part of it, launching his first company, Zip2 Corporation in 1995."

In this sentence, "his' and "he" both refer to Elon Musk, respectively, but their antecedents are not explicitly mentioned. To correctly identify the antecedents of pronouns and other referring expressions in a text is the main task of coreference resolution. Singh et al. (2019) conducted a comprehensive and theoretical investigation of 13 open-source coreference resolution tools and libraries. The study evaluated these tools on various parameters, such as their approach, maintainability, and usage. And a black box testing approach to evaluate these tools' performance was conducted by the author too. Based on the outcomes of their investigation, BART, Berkeley, Reconcile, Stanford, and SpaCy were identified as the most promising tools for further coreference resolution analysis (Singh, 2019).

To enhance the Named Entity Recognition (NER) process, SpaCy will continue to be used as the primary tool in this study to solve co-reference problems. Several python libraries that are related to coreference resolution can be integrated with spaCy, with some popular options being Crosslingual-Coreferences, Neuralcoref, and SpaCy-experimental, etc.

The Crosslingual-Coreference package, originally developed by David Berenstein (2022), is an open-source project from GitHub community (Spacy Universe: Crosslingual coreference, 2023), which operates under the assumption that a trained model with English data and cross-lingual embedding can function for other languages with similar sentence structures. Neuralcoref (Huggingface, 2017/2023) is another open-source project from the GitHub community developed by Huggingface. Neuralcoref is originally based on the fast spaCy parser and utilizes the neural net scoring model described in the paper Deep Reinforcement Learning for Mention-Ranking Coreference Models (EMNLP) by Clark and Manning (2016). Neuralcoref also offers the possibility of training the coreference resolution system on datasets other than English (Clark & Manning, 2016). In our study we will use David Berenstein's Crosslingual-Coreferences package as our coreference resolution for the NER process (David, 2022). Here is the result by using Crosslingual-Coreferences:

Input Text:

"After leaving Penn, Musk headed to Stanford University in California to pursue a PhD in energy physics. However, his move was timed perfectly with the Internet boom, and he dropped out of Stanford after just two days to become a part of it, launching his first company, Zip2 Corporation in 1995. Musk became a U.S. citizen in 2002."

Output Text:

After leaving Penn, Elon Musk moved to Stanford University in California to pursue a PhD in energy physics. However, Elon Musk's move was timed perfectly with the Internet boom, and Elon Musk dropped out of Stanford University in California after just two days to become a part of the Internet boom, launching Elon Musk's first company, Zip2 Corporation in 1995. Elon Musk became a U.S. citizen in 2002.

4.3.3 Relationship Extraction (RE)

Relationship Extraction (RE) is a natural language processing task that aims to extract semantic relationships from text. It involves identifying and extracting relationships between entities of several types, such as people, organizations, and locations, and categorizing them into various semantic categories (Ruder, 2023).

There are several approaches to RE, which include rule-based approaches, supervised approach, and semi-supervised approach (Joshi, 2019). The rule-based approach involves defining a set of rules for the syntax and other grammatical properties of a natural language. These rules are then used to extract information from the text. Although this approach can be effective, it requires significant manual effort to create and can be limited in its ability to handle the complexity and variability of natural language. In a supervised approach, a machine-learning model is trained to detect whether there is a relationship (R) between two entities (E1 and E2) in each sentence (S). The model is trained on a large amount of labeled data and can achieve high accuracy, but it requires a significant amount of labeled data for training. When labeled data is limited, a semi-supervised approach can be utilized. This involves using a set of seed examples (triples) to formulate high-precision patterns that can be used to extract more relations from the text. This approach can be effective in situations where labeled data is scarce, but it may not achieve the same level of accuracy as a fully supervised approach (Joshi, 2019). In our study, we will explore a supervised approach model called REBEL (Relation Extraction by End-to-end Language) to extract the relationships.

REBEL (Relation Extraction by End-to-end Language) Model

The term Relation Extraction has been frequently used in academic literature to refer to a variety of tasks and approaches (Nayak & Ng, 2020). To avoid ambiguity, this study adopts a specific definition of Relation Extraction (RE) as the process of extracting triplets of relations between entities from unprocessed text, without any prior knowledge of entity boundaries, which is commonly known as end-to-end Relation Extraction. In contrast, Relation Classification (RC) pertains to the classification of the relationship between two entities in a specific context.

The REBEL (Relation Extraction by End-to-end Language generation) model, first introduced by Cabot and Navigili (2021), is an autoregressive methodology that presents Relation Extraction as a sequence-to-sequence (seq2seq) challenge, coupled with the REBEL dataset - a sizeable distantly supervised dataset obtained by utilizing a Natural Language Inference (NLI) model. By pre-training an Encoder-Decoder Transformer model using their datasets, Cabot & Navigili's REBEL model attains superior performance compared to various old RE techniques. The REBEL model is accessible in two formats, a standalone model capable of extracting over 200 diverse types of relations, and a pre-trained RE model that can be easily fine-tuned on new RE and RC datasets (Cabot & Navigli, 2021). Further, REBEL is able to integrate with the Python SpaCy library. The present study will employ the REBEL model developed by Cabot and Navigili (2021) as our primary tool for relation extraction. The REBEL output data structure is typically represented in the following format, which is of utmost importance:

{'relation': 'owner of', 'head_span': Elon Musk, 'tail_span': PayPal}

It is essential that the output data structure of Information Extraction is recognizable as input data by the Knowledge Graphs phase, and the REBEL output data structure plays a crucial role in achieving this integration. Thus, the significance of the REBEL output data structure lies in its ability to facilitate a seamless connection with the subsequent stage of knowledge graphs, which ensures a smooth flow of data through the pipeline.

4.3.4 A full Information Extraction Process

At the beginning of this section, we explored the potential use of SpaCy for entity extraction, and discussed challenges that arise in this process, such as the co-reference problem. To address these challenges, we explored David Berenstein's Crosslingual-Coreferences package as our chosen coreference resolution method for NER. Then we explored REBEL for entity and relation Extraction, even though integrating those two components to a stream pipeline is incredibly challenging and time consuming, since they were developed by different individuals and groups. Still, we can make it work, the original components can be found at GitHub, see (Spacy Universe: Crosslingual coreference, 2023) and (Babelscape, 2021).

As a result, the information extraction process utilized in this study is as follows:

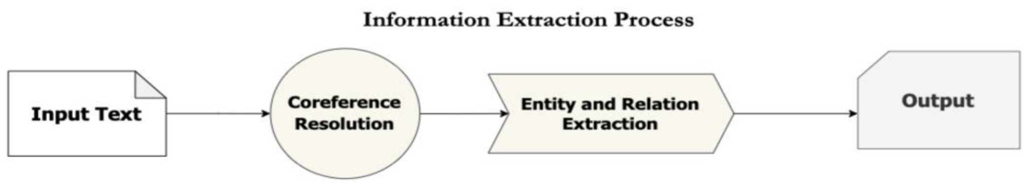


Figure 10. Information Extraction Process

4.4 Artifact: Information Visualization, Knowledge Graph

In the preceding section, we introduced the theory of Knowledge Graphs. A knowledge graph is a directed labeled graph characterized by explicit and well-defined meanings assigned to its labels. Such a graph is composed of nodes, edges, and labels, with any entity, such as individuals, corporations, or even machines, serving as nodes (Stanford edu, 2023). An edge establishes a connection between a pair of nodes and represents a relationship of interest, such as a social or commercial connection between two individuals or organizations. The labels on the edges convey the essence of the relationship, for instance, characterizing the nature of a friendship between two individuals.

The fundamental attributes of a knowledge graph include "unifying data, integrating data sources, and mapping relationships across the data entities" (Sajid, 2022, p.3). Knowledge graph is a computational framework that utilizes machine learning and natural language processing (NLP) to identify and extract entities from unstructured data, facilitating the mapping of relationships between all entities to establish a schema for the graph. One important feature of KG is that the graph reproduces the intricate network of entities and their corresponding relationships, leading to the creation of a simplified data model that allows for a more accessible comprehension of the multifaceted layers of knowledge. The references to the entities and their relationships are then recorded in a graph database, which functions as a knowledge base for knowledge graphs. A typical knowledge graph is a directed graph denoted as G = (N, R), where the set of nodes N represents both entities and literal values. The edges R in turn, indicate the relations among these entities (Jiménez, 2022).

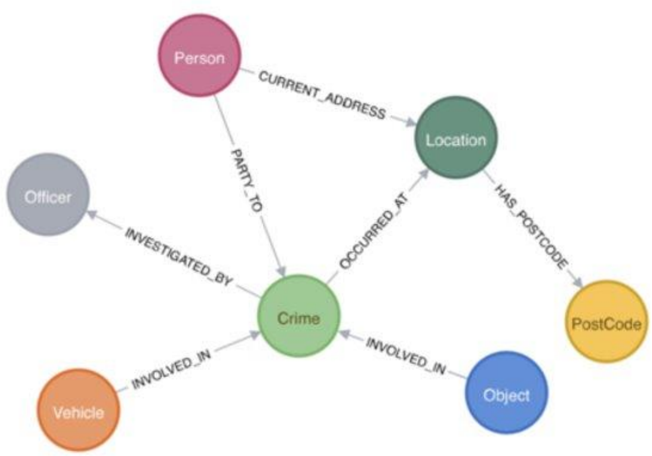


Figure 11. An example of Knowledge Graph Network, Borhade 2021

4.4.1 Advantages of KG

Aside from converting data into machine-interpretable knowledge, a knowledge graph presents several other advantages, such as data unification, ease of availability, visualization of knowledge flow, semantic meaning, and so on (Sajid, 2022). First, data unification, the implementation of a knowledge graph surpasses the fundamental aspects of data assembling and accumulation. It functions as a powerful instrument for knowledge management, facilitating the integration of real-world data and its associated context from a multitude of sources. Whether the data is structured or unstructured, a knowledge graph can unify all data types and act as a sole source of comprehensive knowledge. Second, ease of availability, the Internet contains an ever-expanding pool of information that exceeds our capacity to effectively sift through it. To gain valuable insights from this wealth of data, we require sophisticated tools and systems capable of managing and sorting through vast amounts of information. Knowledge graphs are a particularly useful platform for businesses to disseminate their knowledge across their organizational teams, thereby fostering collaboration and the generation of valuable insights. In addition to integrating data and their relationships within a data layer, a knowledge graph provides a comprehensive representation of real-world data and its intricate interconnections. Moreover, the data network established by a knowledge graph can readily accommodate new data while automatically revealing relevant connections, without necessitating any alterations to the graph structure. Further, the knowledge graph generates a precise graphical illustration of the dissemination of information among data entities in the network. The visualization capability of knowledge graphs renders it an incredibly advantageous tool for tracking workflows, identifying entities that require attention, and detecting patterns over an extended period.

4.2.2 Python NetworkX

There are numerous data visualization tools available for the representation of knowledge graphs. Notably, the R library Igraph and Statnet, as well as the Python library NetworkX, are the most popular ones among many tools. The R library Igraph is a powerful tool for the analysis of graphs and networks, which is built on a core written in C. Igraph is a fast and efficient library that provides a comprehensive set of tools for creating, manipulating, and analyzing graphs (Igraph, R Interface). The Igraph package is widely used in network analysis, social network analysis, and other fields that involve the analysis of complex systems, in which the package includes functions for creating several types of graphs and networks, as well as for computing various properties of graphs, such as centrality measures, clustering coefficients, shortest paths, and so on. It also includes algorithms for community detection, network visualization, and graph layout. Python library NetworkX is another open-source software package for the creation, manipulation, and study of complex networks or graphs (NetworkX, 2023). The library provides tools for building several types of networks, including directed and undirected graphs, multi-graphs, and weighted graphs. NetworkX also offers a range of algorithms and functions for analyzing graphs, such as measuring node centrality, detecting communities, and finding shortest paths between nodes.

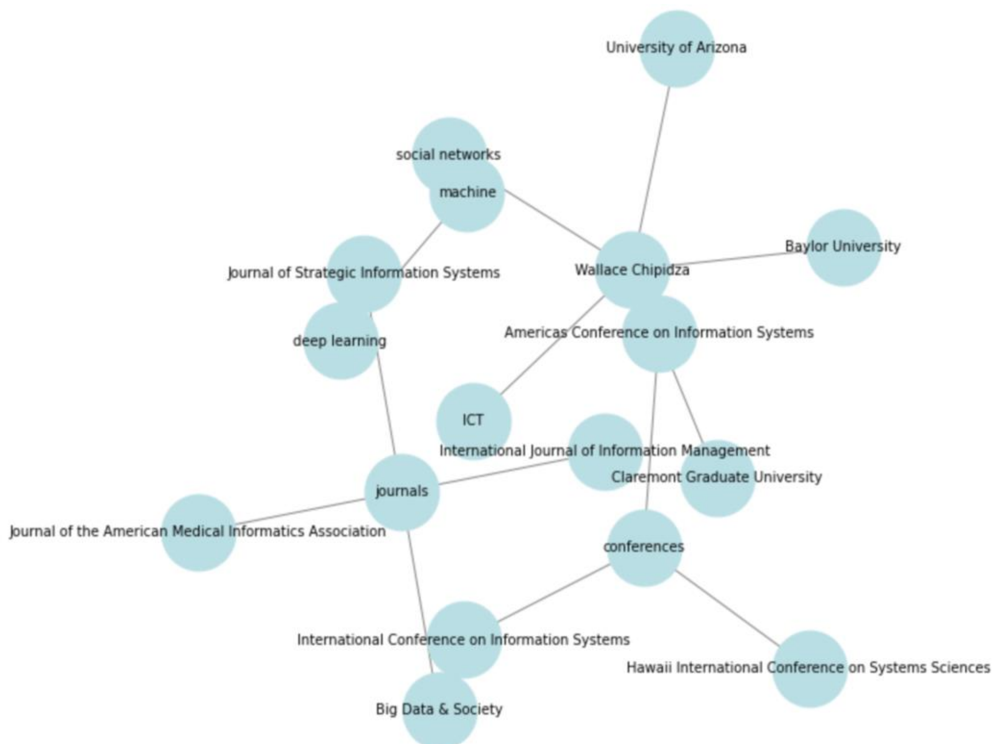


Figure 12. An example of KG drew by Python NetworkX

For this study, Python has been employed as our primary tool for Information Extraction, thus, NetworkX will be used for creating knowledge graph networks. One advantage of NetworkX is it is designed to be easy to use and has a clear and concise API that allows users to efficiently create and manipulate graphs. Another advantage of NetworkX is its flexibility, which permits users to customize graph properties, such as node labels, edge weights, and layout algorithms, to create tailored visualizations of their graphs that meet their specific requirements and preferences.

Moreover, NetworkX offers a range of tools for importing and exporting graph data in various formats, including GML, GraphML, and Pajek, which facilitates the sharing and collaboration of graph data among users. Additionally, NetworkX has an active and thriving open-source community with a large and expanding user base, resulting in a wealth of resources such as documentation, tutorials, and user forums, that are available for learning and utilizing the library.

4.5 Our Delivery: T-3C Model

Thus, based on design science activities, we designed and developed a comprehensive automated information-processing model that can streamline the entire information processing cycle, from information search and collection to extraction and visualization. The components or artifacts of our model are composed of advanced and innovative information technologies such as web scraping, natural language processing and Information Extraction, and knowledge graphs. By doing so, we aim to streamline the knowledge graph creation process and make it more accessible to a broader range of users.

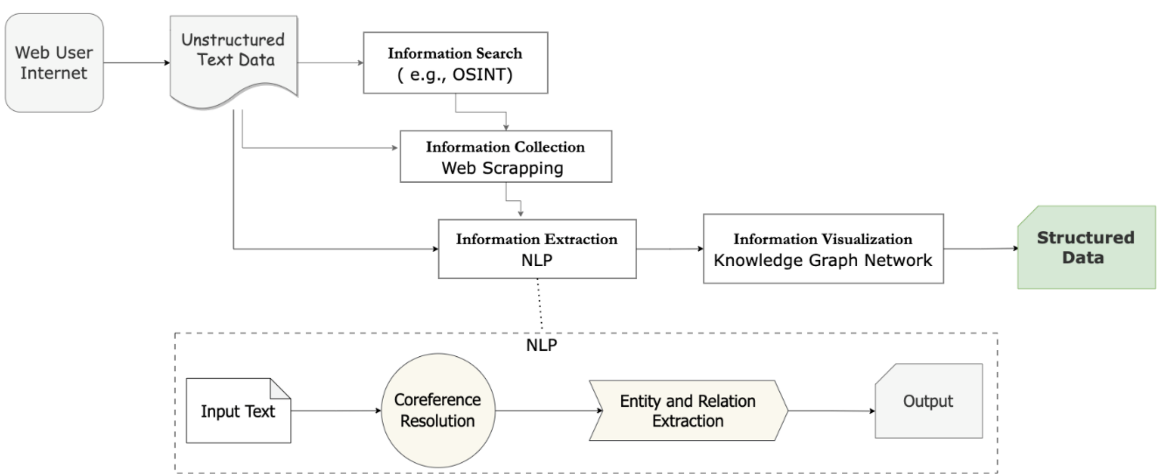


Figure 13. Improved model with IT artifact, T-3C Model

In Chapter 3, we had proposed our conceptual model (see Fig 5), the model aims to streamline the automatic information flow process from information search to information collection, to information extraction, and to information visualization. Now, we add IT components to each stage, and we have designated our proposed model Automated Information Process and Visualization as T-3C Model, in which, T-3C represents Thullen – Chipidza, Chatterjee, Chinazunma.

One distinguishing feature of our model is that it enables the information flow to initiate directly from unstructured data to web scraping or NLP process, without necessarily an information search process since in some conditions, the process can start from existing documents directly. The need for such processes would depend on specific real-life scenarios. In the following chapter, we will perform use case studies and apply varying scenarios to our model.

Our goal is by using our model users can collect valuable information as well as to extract high-quality information more effectively and efficiently in the information explosion age. Further, by transforming unstructured textual data into a structured knowledge-based data model, users can visualize textual information and present key information in a more effective way.

In the following section, several real-world case studies will be presented to evaluate the practicality and feasibility of the proposed model.

Chapter 5. Model Demonstration: Case Study

In an earlier chapter, we highlighted the significance importance of demonstration in the Design Science Research process. In this section, we will provide multiple case studies of the model, which will present an opportunity to confirm the practical effectiveness of the framework artifacts. And the demonstration includes adapting the model to a real-world use case study that enables researchers to evaluate the model's capacity to accomplish various research objectives. One of our model's critical objectives is to enhance the efficiency of the information gathering process. The utilization of OSNIT and Web Scraping is among the techniques employed to attain this goal. Another significant objective of the model is to facilitate greater comprehension of the Information Extraction and Visualization process, especially converting unstructured textual information to structured KG data model. This feature can also be demonstrated through the case study.

5.1 System Artifact with Selected IT Tools

Before we start to dem the model, we would like to discuss the challenges we encountered during our model building process and the obstacles we have to overcome in the model implementation process first.

In the conceptual level of model design, we had stated that machine-learning algorithms can only process numerical data, which means that textual data cannot be directly inputted into these algorithms. As a result, textual data needs to be transformed into a numerical format before it can be processed by these algorithms (Shankar, 2022). During our case study, we encountered significant challenges when attempting to apply to our information process model. These challenges included difficulties on web scraping during the information collection process, challenges arising from the information extraction process, and challenges related to information visualization. Additionally, there were challenges associated with building an automatic information pipeline that could allow for seamless information flow throughout the entire model.

The most significant and critical challenge we faced was ensuring that information flowed through the model smoothly and seamlessly from beginning to end. The first obstacle in achieving this goal was the selection of appropriate methods and tools for each process involved in information gathering, extraction, and visualization. Numerous tools exist for each of these processes. For example, in the case of information extraction, the Stanford NLP tool is a popular option for Java developers. However, selecting a specific tool also means committing to using a particular programming language throughout the entire information flow process, or potentially encountering compatibility issues if different programming languages are used. The second challenge was the seamless integration of all the tools. Although we used Python and its extended library as our primary tools for information flow in our case study, integrating the various extension Python components was a significant challenge. This was because these packages are open source and developed by different individuals or groups and may or may not be compatible with one another. Finding the right tools and making them work seamlessly together to create an automated information flow was indeed challenging.

One of the other most significant challenges we faced was ensuring the accuracy of our model. The final output of the model was highly dependent on the information extraction process's effectiveness, particularly with respect to the entities and relationships extraction. The extraction of information from the text was performed using the REBEL tool, which relied on pre-trained datasets for relationship extraction. However, to enhance the accuracy of the model, it was essential to augment the training data for the extraction process. This poses a daunting challenge, as it requires enormous amounts of high-quality data and requires the ability to train the model to improve the preciseness of the model.

Another challenge pertained to the visualization of information, specifically the graph network mapping. When the input text was too long, the information extraction process could be complex, especially the graph network tended to become overcrowded. To address this issue, we can filter the graph network based on keywords, and NetworkX will only present the nodes and edges we are interested in. For example, in case study 5.2 Elon Musk, the graph network is crowded (see Figure 25), after we filter keywords Elon Musk, the graph will only present key information about Elon Musk (see Figure 26).

Another significant challenge is the web scraping process, which needs to inspect through website HTML to extract relevant information. This process can be time-consuming and requires many resources. The differences in formatting and templates can also result in inconsistencies in the extracted data, which requires a certain amount of time to clean up the data and prepare for next stage use.

Lastly but not the most important challenge is using OSINT for the information search process. OSINT is particularly useful for information search as well as a valuable source of information that can provide insights into various areas of interest in information overload. However, despite its many advantages, we found some challenges and limitations associated with the collection and analysis of OSINT data during our research. One of the main challenges of OSINT is the sheer volume of data produced during the collection process. Collecting OSINT from various sources, such as social media, news outlets, and blogs, can generate a massive amount of data that needs to be analyzed to derive useful insights. While there are automated tools available to filter and process the data, the volume of data remains a significant challenge for the OSINT gatherer. Even with advanced AI (Artificial Intelligence) tools, the processing of enormous amounts of data can be time-consuming and expensive. Another challenge of OSINT is the time-consuming process of data cleaning. After collecting the data, humans need to view the output of automated tools to ensure that the collected data is reliable and trustworthy. Furthermore, the data needs to be compared with classified data, especially in the context of military and commercial information, to verify its reliability and relevance. This process can be time-consuming and requires significant resources, particularly when dealing with large volumes of data. The reliability of OSINT sources is another significant challenge. When OSINT is used in intelligence contexts, sources must be thoroughly verified by classified sources before they can be trusted. Governments may broadcast inaccurate information to mislead the OSINT-gathering process. Using unreliable sources can lead to incorrect conclusions and potentially dangerous situations, particularly in national security. Nevertheless, with the use of advanced AI tools and careful data analysis, OSINT can continue to be a valuable information search and information retrieval tool.

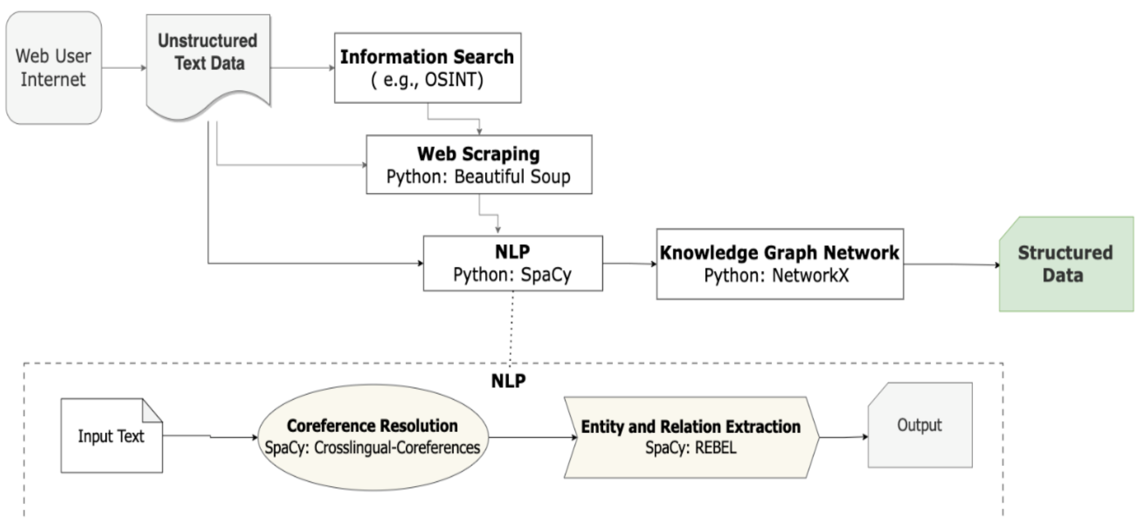


Figure 14. T-3C Model with Python Libraries

Despite encountering those challenges, we managed to overcome them and applied our model to real-world scenarios successfully. In our prior research, we had delineated the methodology and tools that will be deployed to each stage of the IT artifact, specifically, Python and its associated libraries have been chosen as the principal instruments to explain the practical application of our study. The model now includes example Python packages, which will apply to our real-world use case, a completed model with IT tools as Figure 14.

In summary, the demonstration of a framework is crucial for validating its practical effectiveness. It provides an opportunity to adapt the artifact to a real-world use case study, which enables the researchers to validate the framework's practical effectiveness. The use case study presented in this section aims to enhance the efficiency of the information gathering process and facilitate greater comprehension of the collected data through the utilization of Web Scraping, Information Extraction, and Knowledge Graph techniques. Therefore, in the rest of this section, we will present three case studies to evaluate the efficacy of our framework.

5.2 Case Study 1: CGU Faculties

Scenarios: A student is considering applying for the Information and Technology doctoral program at Claremont Graduate University (CGU). The student desires to obtain comprehensive background information on each faculty member in this department, including the identities of the professors, their respective research areas, publications, work experiences, and so on.

Phase 1. Information Search

Sometimes, we might not need to conduct a complicated information search process. In this case, through a simple keywords search CGU Information and Technology Faulty, we found the CGU Center for Information & Technology (CISAT) department faculties:

https://www.cgu.edu/school/center-for-information-systems-and-technology/faculty/

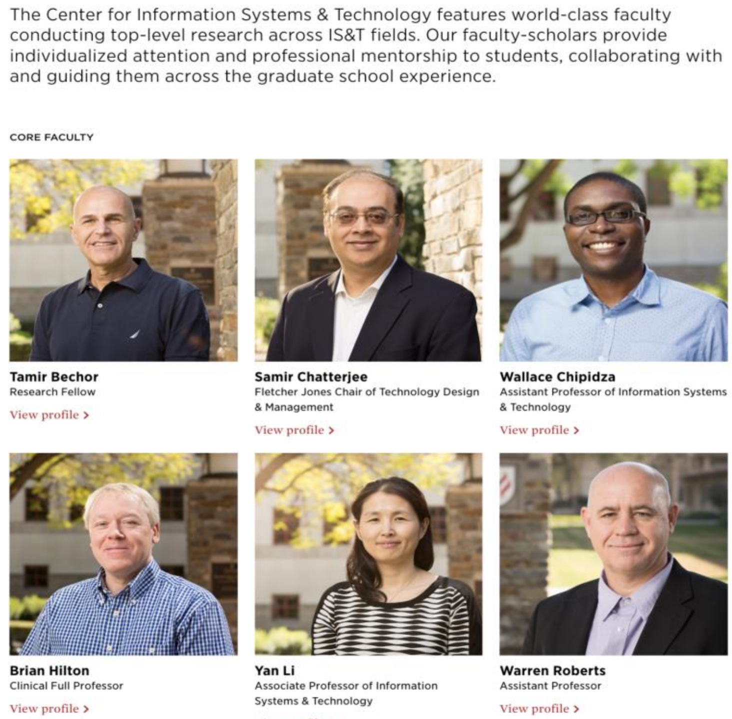


Figure 15. CGU CISAT Faculties Websites

For each faulty the detail information as below:

Figure 16. an example of CGU CISAT Faculty, Dr. Wallace Chipidza

Phase 2. Information Collection through Web Scraping

Through the Python library Beautiful Soup, the key information for all the faculties can be collected and saved to local devices. Web Scraping information as below:

```
Tamir Bechor
Research Fellow
https://www.cgu.edu/people/tamir-bechor/
Tamir Bechor is a research fellow in Claremont Graduate University¹s Center for Information Systems and Technology (CISAT). He also teaches in CGU’s Tran
Bechor has extensive knowledge in the field of cybersecurity. His knowledge and experience in handling national security problems contributes to his capa
Bechor received his PhD from Tel Aviv University in Israel before establishing his career at CGU. Bechor is a permanent member of the advisory board for

Samir Chatterjee
Fletcher Jones Chair of Technology Design & Management
https://www.cgu.edu/people/samir-chatterjee/
Samir Chatterjee is the Fletcher Jones Chair of Technology Design & Management at CGU’s Center for Information Systems & Technology (CISAT). He is also c
Chatterjee received a Bachelor’s of Technology in Electronics & Telecommunications Engineering (1988) from Jadavpur University, India; as  well as an MS
Throughout his career, Chatterjee has had substantial influence on the advancement of technology within the health care industry. He founded the Network
Since 2006 he has been an evangelist and champion of design science as a research method in the IS community. He started the successful DESRIST series of
His current projects include designing ICT and mobile solutions to address management of such chronic diseases as obesity/diabetes as well as oral health

Wallace Chipidza
Assistant Professor of Information Systems & Technology
https://www.cgu.edu/people/wallace-chipidza/
Wallace Chipidza is an assistant professor in the Center for Information Systems & Technology at Claremont Graduate University. He holds a PhD in informa
Chipidza mainly researches how and why social networks change over time, the impacts of those changes, and the interventions that moderate those impacts.
Chipidza’s work has been published in multiple prestigious journals including Journal of Strategic Information Systems, Decision Support Systems, Journal
```

Figure 17. an extract of CGU CISAT Faculties Information been collected through Python BeautifulSoup

The information collected for Dr. Wallace Chipidza by Web Scraping:

Wallace Chipidza

Assistant Professor of Information Systems & Technology https://www.cgu.edu/people/wallace-chipidza/

Wallace Chipidza is an assistant professor in the Center for Information Systems & Technology at Claremont Graduate University. He holds a PhD in information systems from Baylor University and an MS in computer science from the University of Arizona. Chipidza research how and why social networks change over time, the impacts of those changes, and the interventions that moderate those impacts. His other interest lies in designing ICT based solutions to problems afflicting vulnerable populations in developing countries. To answer his research questions, Chipidza uses a variety of advanced techniques including machine and deep learning, network modeling, and simulations. Chipidza’s work has been published in multiple prestigious journals including Journal of Strategic Information Systems, Decision Support Systems, Journal of the American Medical Informatics Association, Big Data & Society, and International Journal of Information Management. His work has been at the International Conference on Information Systems, the Hawaii International Conference on Systems Sciences, and the Americas Conference on Information Systems, among other conferences.

Phase 3. Information Extraction Process

Our information extraction methodology will be continually applied to Dr. Chipidza as a representative case.

Input text:

Wallace Chipidza is an assistant professor in the Center for Information Systems & Technology at Claremont Graduate University. He holds a PhD in information systems from Baylor University and an MS in computer science from the University of Arizona. Chipidza research how and why social networks change over time, the impacts of those changes, and the interventions that moderate those impacts. His other interest lies in designing ICT based solutions to problems afflicting vulnerable populations in developing countries. To answer his research questions, Chipidza uses a variety of advanced techniques including machine and deep learning, network modeling, and simulations. Chipidza's work has been published in multiple prestigious journals including Journal of Strategic Information Systems, Decision Support Systems, and Journal of the American Medical Informatics Association, Big Data & Society, and International Journal of Information Management. His work has been at the International Conference on Information Systems, the Hawaii International Conference on Systems Sciences, and the Americas Conference on Information Systems, among other conferences.

Coreference Resolution

In the preceding section, the coreference problem in the entity extraction process was addressed, wherein the original text's entity extraction capability was found to be inadequate in extracting critical entity information such as pronouns (e.g., "he," "his," etc.). The Python library Neuralcoref was also introduced to address the issue, and the outcomes of its application are presented herewith. Coreference Solution Result (Python Crosslingual-Coreferences):

Output Text:

Wallace Chipidza is an assistant professor in the Center for Information Systems & Technology at Claremont Graduate University. Wallace Chipidza holds a PhD in information systems from Baylor University and an MS in computer science from the University of Arizona.

Wallace Chipidza researched how and why social networks change over time, the impacts of those changes, and the interventions that moderate those changes. Wallace Chipidza's other interest lies in designing ICT based solutions to problems afflicting vulnerable populations in developing countries. To answer Wallace Chipidza's research questions, Wallace Chipidza uses a variety of advanced techniques including machine and deep learning, network modeling, and simulations.

Wallace Chipidza's work has been published in multiple prestigious journals including Journal of Strategic Information Systems, Decision Support Systems, and Journal of the American Medical Informatics Association, Big Data & Society, and International Journal of Information Management. Wallace Chipidza's work has been included in the proceedings of the International Conference on Information Systems, the Hawaii International Conference on Systems Sciences, and the Americas Conference on Information Systems, among other conferences.

Entities and Relationship Extraction (Python REBEL):

Extract Entities and Relations from above paragraphs through Python REBEL, 14 relationships had been extracted:

```
(0, 15): {'relation': 'educated at', 'head_span': Wallace Chipidza, 'tail_span': Claremont Graduate University}
(0, 28): {'relation': 'educated at', 'head_span': Wallace Chipidza, 'tail_span': Baylor University}
(0, 38): {'relation': 'educated at', 'head_span': Wallace Chipidza, 'tail_span': University of Arizona}
(0, 50): {'relation': 'field of work', 'head_span': Wallace Chipidza, 'tail_span': social networks}
(0, 79): {'relation': 'field of work', 'head_span': Wallace Chipidza, 'tail_span': ICT}
(108, 110): {'relation': 'has part', 'head_span': machine, 'tail_span': deep learning}
(110, 108): {'relation': 'part of', 'head_span': deep learning, 'tail_span': machine}
(132, 130): {'relation': 'instance of', 'head_span': Journal of Strategic Information Systems, 'tail_span': journals}
(142, 130): {'relation': 'instance of', 'head_span': Journal of the American Medical Informatics Association, 'tail_span': journals}
(150, 130): {'relation': 'instance of', 'head_span': Big Data & Society, 'tail_span': journals}
(156, 130): {'relation': 'instance of', 'head_span': International Journal of Information Management, 'tail_span': journals}
(174, 198): {'relation': 'instance of', 'head_span': International Conference on Information Systems, 'tail_span': conferences}
(181, 198): {'relation': 'instance of', 'head_span': Hawaii International Conference on Systems Sciences, 'tail_span': conferences}
(190, 198): {'relation': 'instance of', 'head_span': Americas Conference on Information Systems, 'tail_span': conferences}
```

Figure 18. Entity and Relation Extraction by Python REBEL

Phase 4. Information Visualization- Knowledge Graph

From REBEL data model results, we can map the relationships between entities by Python NetworkX:

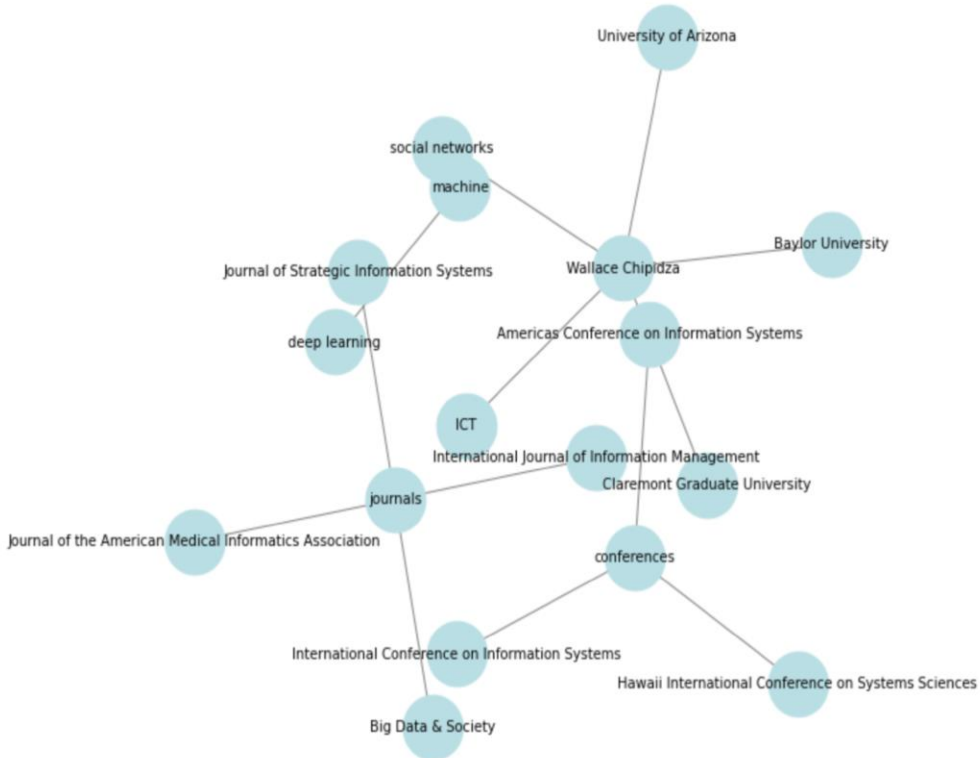


Figure 19. Knowledge Graph Model for Dr. Wallace Chipidza

From the above case, we successfully collected Dr. Chipidza profiled information from the Internet and transformed it into a Knowledge Graph data model.

Here is the result another CGU faculty member, Dr. Samir Chatterjee:

Input text:

Samir Chatterjee is the Fletcher Jones Chair of Technology Design & Management at CGU's Center for Information Systems & Technology (CISAT). He is also considered a leading technology designer and strategist for 21st-century health care. His entry into the healthcare field has been via Telemedicine. Today he leads the emerging field of Persuasive Technology, a stimulating interdisciplinary research field that focuses on how interactive technologies and services can be designed to influence people's attitudes and support positive behavior change.

Chatterjee received a Bachelor of Technology in Electronics & Telecommunications Engineering (1988) from Jadavpur University, India; as well as an MS (1991) and PhD (1994) in Computer Science from the School of Computer Science, University of Central Florida. He joined CGU in July 2001. In May 2015, he was awarded the distinguished lifetime achievement award for contributions to Design Science Research, presented by the IS design community. He is also an adjunct faculty at Keck Graduate Institute, where he teaches a course on Healthcare Informatics. He has also taught in the Drucker School of Management's Executive MBA Program.

Output Graph as below:

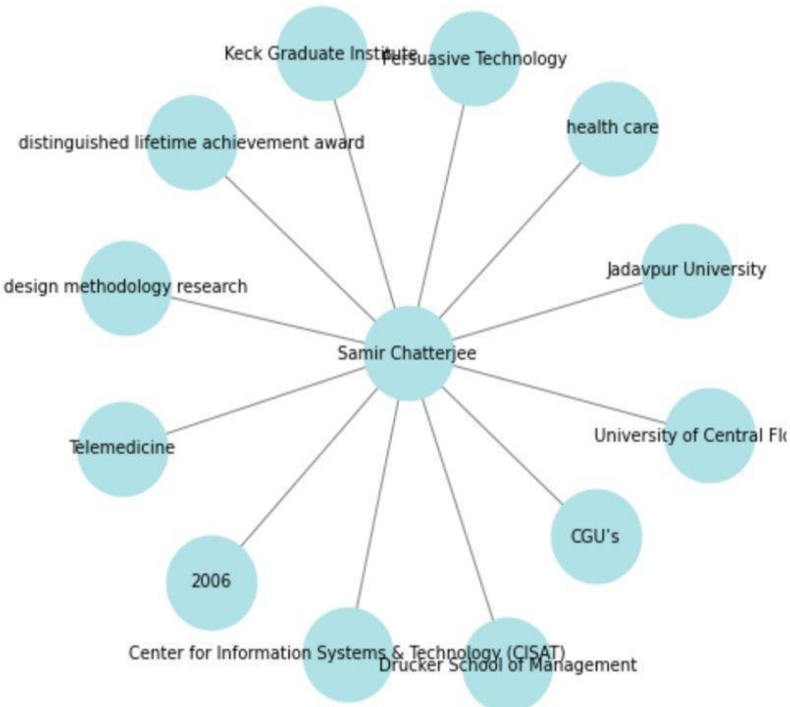


Figure 20. Knowledge Graph Model for Dr. Samir Chatterjee

5.3 Case Study 2: Elon Musk Biography

Let us take Elon Musk as another example; we would like to know some essential information and events about Elon Musk.

A traditional way would be to search for Elon Musk from Google. There are about 656,000,000 results about Elon Musk, which are from news, social media, videos, and so on. The challenges are how could we find useful information we need and how could we transform that large scale of textual information to structured knowledge? Unstructured Raw Information about Elon Musk, *656,000,000* results:

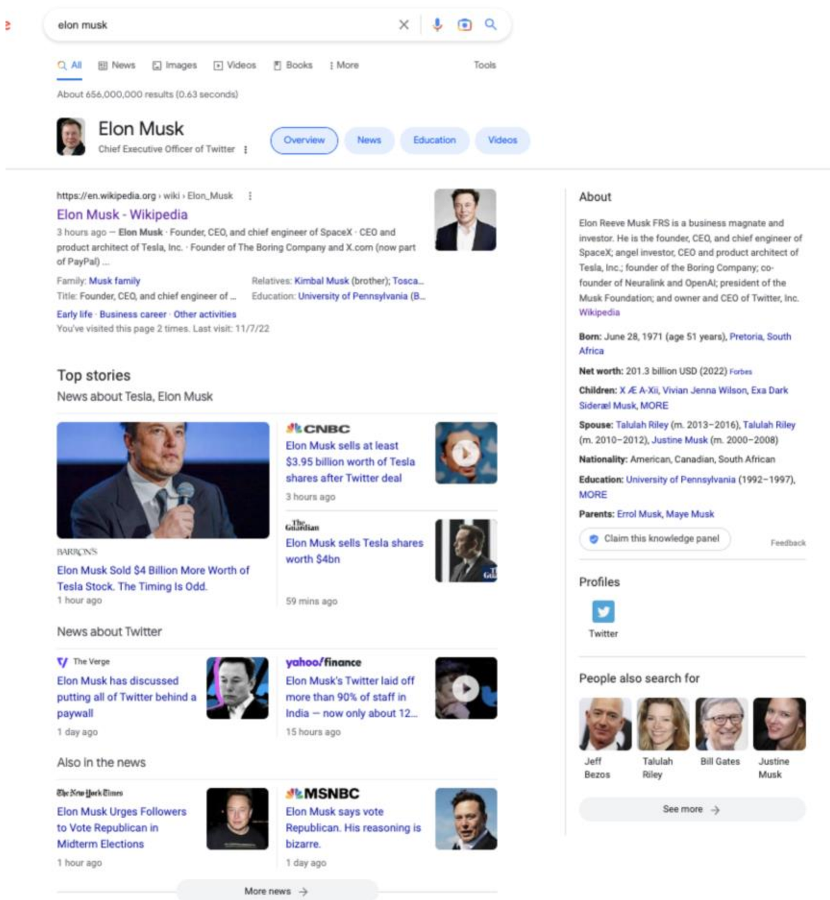


Figure 21. Google Search Result for Elon Musk, May 2023

Based on our model, we will do the following steps - information search and collection, information extraction, and information visualization.

Information Search and Web Scraping

Through the principles and methods of Open-Source Intelligence, we can find some reliable resources about Elon Musk. We can scrap as many as websites if we need, but let us look at *https://www.biography.com/business-figure/elon-musk*

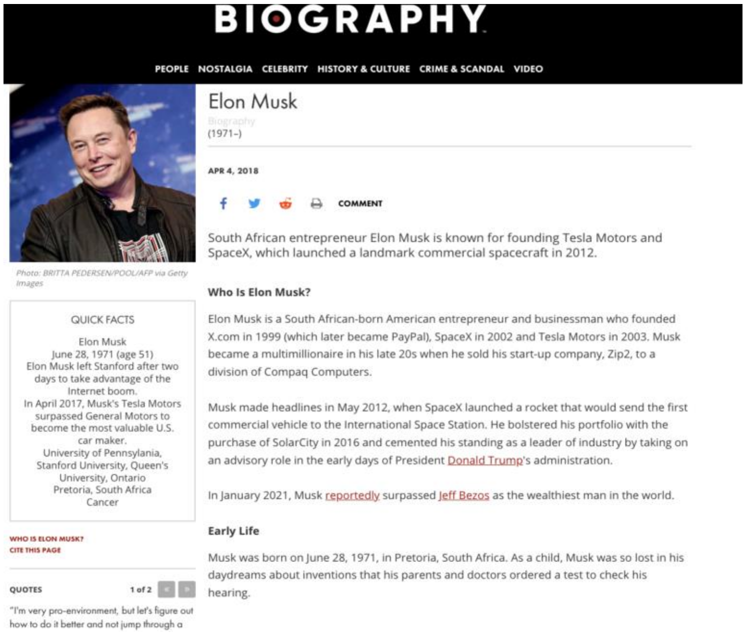


Figure 22. Information Search: An Information Repository for Elon Musk

First, scraping useful information from Figure 22. We used Python Library BeautifulSoup to scrape the information. Scraping result as below:

```
[41]: [[ Elon Musk is a South African-born American entrepreneur and businessman who founded X.com in 1999 (which later became PayP
      al), SpaceX in 2002 and Tesla Motors in 2003.],
       [Musk became a multimillionaire in his late 20s when he sold his start-up company, Zip2, to a division of Compaq Computer
      s.],
       [  Musk made headlines in May 2012, when SpaceX launched a rocket that would send the first commercial vehicle to the Intern
      ational Space Station.],
       [He bolstered his portfolio with the purchase of SolarCity in 2016 and cemented his standing as a leader of industry by taki
      ng on an advisory role in the early days of President Donald Trump's administration.],
       [In January 2021, Musk reportedly surpassed Jeff Bezos as the wealthiest man in the world.],
       [Musk was born on June 28, 1971, in Pretoria, South Africa.],
       [As a child, Musk was so lost in his daydreams about inventions that his parents and doctors ordered a test to check his hea
      ring.],
       [At about the time of his parents' divorce, when he was 10, Musk developed an interest in computers.],
       [He taught himself how to program, and when he was 12 he sold his first software: a game he created called Blastar.],
       [In grade school, Musk was short, introverted and bookish.],
       [He was bullied until he was 15 and went through a growth spurt and learned how to defend himself with karate and wrestlin
      g.],
       [Musk's mother, Maye Musk, is a Canadian model and the oldest woman to star in a Covergirl campaign.],
       [When Musk was growing up, she worked five jobs at one point to support her family.],
       [Musk's father, Errol Musk, is a wealthy South African engineer.],
       [ Musk spent his early childhood with his brother Kimbal and sister Tosca in South Africa.],
       [His parents divorced when he was 10.],
       [At age 17, in 1989, Musk moved to Canada to attend Queen's University and avoid mandatory service in the South African mili
      tary.],
       [Musk obtained his Canadian citizenship that year, in part because he felt it would be easier to obtain American citizenship
      via that path.],
       [In 1992, Musk left Canada to study business and physics at the University of Pennsylvania.],
       [He graduated with an undergraduate degree in economics and stayed for a second bachelor's degree in physics.],
       [After leaving Penn, Musk headed to Stanford University in California to pursue a PhD in energy physics.],
       [However, his move was timed perfectly with the Internet boom, and he dropped out of Stanford after just two days to become
      a part of it, launching his first company, Zip2 Corporation in 1995.],
       [Musk became a U.S. citizen in 2002.],
```

Figure 23. Information Collection: Web Scraping result about Elon Musk

Information Extraction: Named Entity Recognition and Relation Extraction

Entity and Relationship Extraction Results based on REBEL:

```
(22, 28): {'relation': 'employer', 'head_span': Elon Musk, 'tail_span': Tesla Motors}
(22, 31): {'relation': 'employer', 'head_span': Elon Musk, 'tail_span': SpaceX}
(28, 22): {'relation': 'founded by', 'head_span': Tesla Motors, 'tail_span': Elon Musk}
(31, 22): {'relation': 'founded by', 'head_span': SpaceX, 'tail_span': Elon Musk}
(22, 69): {'relation': 'owner of', 'head_span': Elon Musk, 'tail_span': PayPal}
(69, 22): {'relation': 'founded by', 'head_span': PayPal, 'tail_span': Elon Musk}
(22, 116): {'relation': 'owner of', 'head_span': Elon Musk, 'tail_span': Zip2}
(116, 22): {'relation': 'owned by', 'head_span': Zip2, 'tail_span': Elon Musk}
(172, 22): {'relation': 'owned by', 'head_span': SolarCity, 'tail_span': Elon Musk}
(22, 172): {'relation': 'owner of', 'head_span': Elon Musk, 'tail_span': SolarCity}
(22, 221): {'relation': 'said to be the same as', 'head_span': Elon Musk, 'tail_span': wealthiest man in the world}
(217, 221): {'relation': 'said to be the same as', 'head_span': Jeff Bezos, 'tail_span': wealthiest man in the world}
(221, 22): {'relation': 'said to be the same as', 'head_span': wealthiest man in the world, 'tail_span': Elon Musk}
(221, 217): {'relation': 'said to be the same as', 'head_span': wealthiest man in the world, 'tail_span': Jeff Bezos}
(22, 235): {'relation': 'date of birth', 'head_span': Elon Musk, 'tail_span': June 28, 1971}
(22, 241): {'relation': 'place of birth', 'head_span': Elon Musk, 'tail_span': Pretoria}
(22, 243): {'relation': 'country of citizenship', 'head_span': Elon Musk, 'tail_span': South Africa}
(241, 243): {'relation': 'country', 'head_span': Pretoria, 'tail_span': South Africa}
(243, 241): {'relation': 'capital', 'head_span': South Africa, 'tail_span': Pretoria}
(374, 22): {'relation': 'designed by', 'head_span': Blastar, 'tail_span': Elon Musk}
(22, 436): {'relation': 'mother', 'head_span': Elon Musk, 'tail_span': Maye Musk}
(436, 22): {'relation': 'child', 'head_span': Maye Musk, 'tail_span': Elon Musk}
(22, 497): {'relation': 'father', 'head_span': Elon Musk, 'tail_span': Errol Musk}
(497, 22): {'relation': 'child', 'head_span': Errol Musk, 'tail_span': Elon Musk}
(22, 530): {'relation': 'sibling', 'head_span': Elon Musk, 'tail_span': Kimbal}
(22, 533): {'relation': 'sibling', 'head_span': Elon Musk, 'tail_span': Tosca}
```

Figure 24. Information Extraction: Entity and Relationship Extraction from Elon Musk Textual Data

Information Visualization: knowledge Graph

For information visualization, we used the python library NetworkX to map the graphs. The full entity and relationship results are as follows:

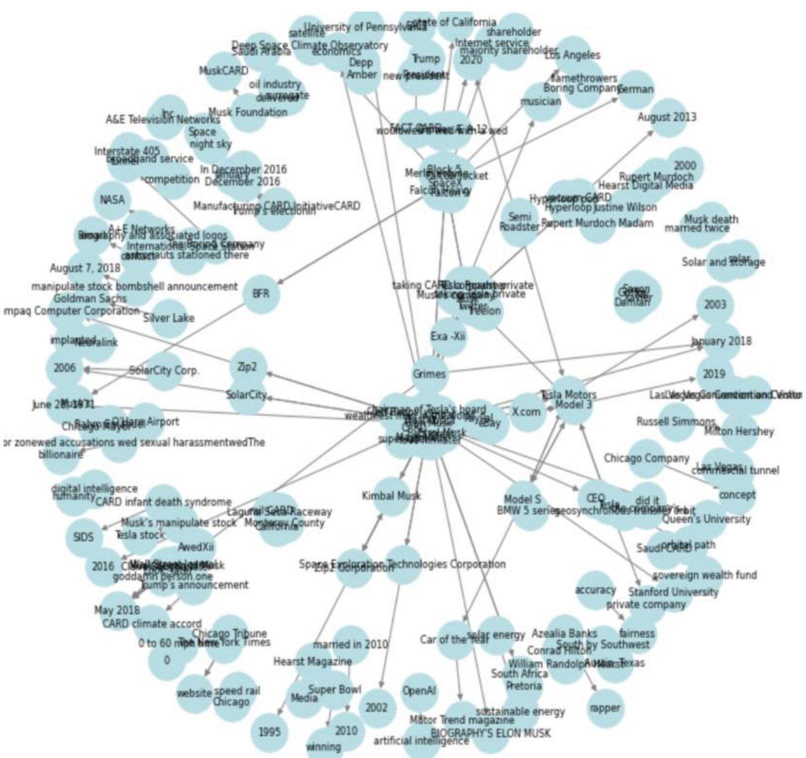


Figure 25. Information Visualization: Knowledge Graph about Elon Musk Network

The layout can be adjusted and filtered based on keywords. Filtered irrelevant information, and only keep key information, new graph as below:

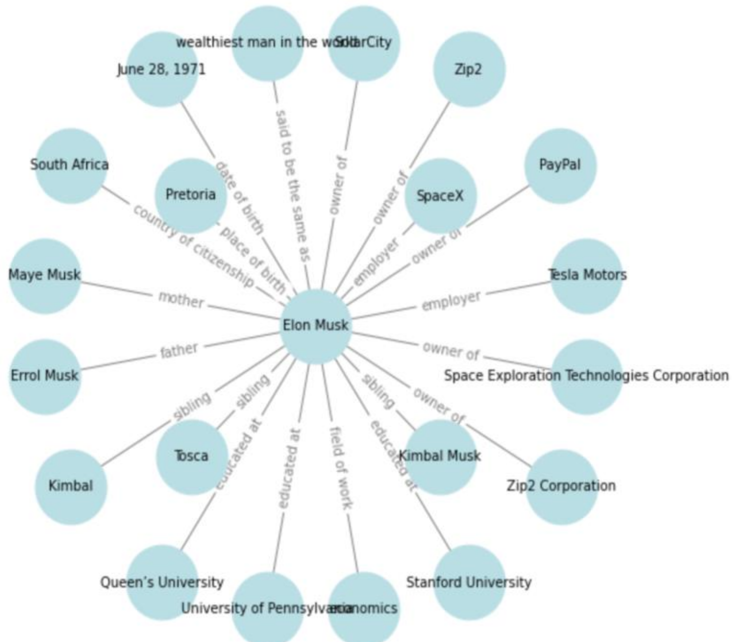


Figure 26. Information Visualization: Key information about Elon Musk

5.4 Model Extension: ChatGPT and T-3C Model

ChatGPT is a large language model trained by OpenAI and based on the GPT-3.5 architecture. It is one of the most advanced and powerful natural language processing models currently available, with a range of potential applications in fields such as machine translation, text summarization, and conversational agents (*Introducing Chatgpt*, 2023).

ChatGPT offers a valuable solution for information collection through the facilitation of natural language queries. One of the primary advantages of utilizing ChatGPT for this study is its inherent ability to process and analyze natural language. As a result, individuals without technical proficiency or expertise in data analytics may still effectively extract relevant and useful information by simply providing queries in natural language, leaving ChatGPT to perform the analytical work. Additionally, ChatGPT possesses prominent levels of scalability by enabling analysis of large datasets. This capacity enables ChatGPT to perform data processing and analysis at impressive speeds and with remarkable accuracy, thus allowing users to collect information from millions of data points in mere minutes.

Accordingly, ChatGPT represents a viable tool for conducting information searches and collecting relevant data. The resulting output generated by ChatGPT can subsequently serve as an input text for conducting entity and relationship extraction, as well as data visualization. An illustration of ChatGPT's application within the context of the T-3C model is presented below. At ChatGPT online interface, we queried some information about Michael Jackson:

Who is Michael Jackson? By ChatGPT

Michael Jackson (1958-2009) was an American singer, songwriter, dancer, and actor, who is widely regarded as one of the most influential musicians of the 20th century. He started his career as a child in the group "The Jackson 5" alongside his brothers. As a solo artist, he released many successful albums, including "Thriller," which is the best-selling album of all time. He was also known for his unique dance moves, such as the Moonwalk, and for his elaborate music videos.

Throughout his career, Jackson won numerous awards, including multiple Grammy Awards, and he is credited with revolutionizing the music industry with his innovative style and techniques. Jackson was also involved in humanitarian work, including founding the "Heal the World Foundation" to help children around the world.

However, Jackson's personal life was also the subject of controversy and allegations of child sexual abuse, which he denied. Jackson died in 2009 at the age of 50 from cardiac arrest.

Applied the textual information to T-3C Model information visualization stage:

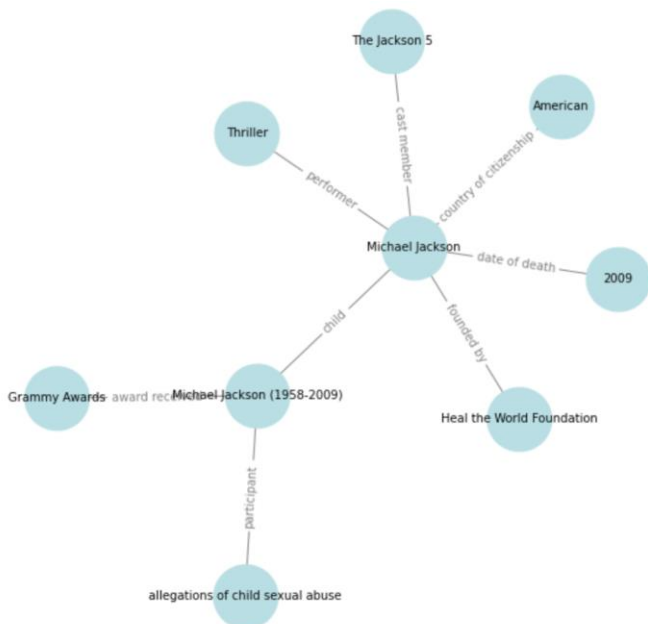


Figure 27. An example of ChatGPT Applied to T-3C Model.

Challenges with ChatGPT

Despite its many strengths, there are also some problems with ChatGPT. One of the fundamental issues is the quality of the data. ChatGPT relies on enormous amounts of unstructured data to provide accurate results. If the training data is incomplete or biased, it can affect the quality of the information extracted - which can result in the model producing biased or discriminatory output. Additionally, language models like ChatGPT are only capable of generating text based on the patterns they have learned from the input data, which means that they are not able to reason or understand concepts in the same way that humans do (Darling, 2023). ChatGPT attempts to generate accurate responses, however, its reliance on patterns rather than verified facts and data often results in inaccuracies.

Thus, ChatGPT faces information reliability issues during information collection, and it should not be considered as a credible source for information collection process and should not be cited in academic research as well. Therefore, users should exercise cautious when relying solely on ChatGPT for the information collection process and consider still using web scraping to ensure the accuracy and reliability of the information they gather.

Chapter 6. Model Evaluation

The evaluation process in DSR involves testing the artifact created in real-world situations to assess its effectiveness in addressing proposed problems. The evaluation is a critical stage in the DSR process as it helps to identify any potential issues with the artifact and to ensure that it is fit for purpose. In the last chapter, we managed to overcome all the challenges and applied our model to real-world scenarios successfully. Our next step involves assessing the efficiency and effectiveness of our model based on the use case studies we conducted.

6.1 DSR Evaluation Process

Design science research focuses on creating and evaluating innovative solutions to real-world problems. It involves the creation of technical artifacts, such as systems or software, and their evaluation based on both technical and social factors. The evolution of design science research has led to the development of evaluation methods that consider both the technical performance of artifacts and their impact on the socio-technical system or organizational context.

The evaluation of the artifacts will be divided into two parts. The first part will focus on the technical performance of the artifact and will involve measurements such as efficiency, utility, performance, usability, innovation, acceptance, and so on. The approaches used to evaluate the technical performance of the artifact will include analytical modeling, simulation, function testing, integer programming, experimental design, formal verification, and unit testing, etc. These methods are designed to provide a comprehensive understanding of the technical capabilities of the artifact and how it performs under different conditions. The second part of the evaluation will focus on the socio-technical system or organizational impact of the artifact. The measurements used in this part of the evaluation will include efficiency, improvements, and statistical evidence of significant gains. The approaches used to evaluate the socio-technical system or organizational impact will include quantitative surveys, qualitative interviews, field studies through deployment, case studies, and controlled studies, etc. These methods are designed to provide insights into how the artifact affects the socio-technical system or organizational context in which it is deployed.

In our study, we will discuss both technical evaluation about the model itself, and at the same time, we will invite volunteers to join our experimental survey and interviews to conduct social technical impacts on the model as well. We will use our framework to run some topics they are interested in and then show them the results. The volunteers will provide their insights and feedback about the model. This feedback can be used to identify any issues with the artifact and to refine it to better address the intended problem.

6.2 Technical Performance Evaluation

We discussed the challenges and obstacles during our model design and implementation process. It was a painful process, but once the model and tools are set up, it can be extremely efficient and time saving. Consider the first use case of CGU Faculties as an illustration, whereby the quantitative assessment of the model will be based on two key metrics: efficiency and accuracy.

Suppose a user intends to obtain information on all the faculties of CGU CISAT from the internet. This activity involves opening the webpage of each individual faculty, reading the information, and either saving the textual information to local storage or taking notes. To complete the entire process, each faculty may take between 15 to 20 minutes. Given that there are nine primary faculties at CGU CISAT, the total time consumption for collecting information on all of them would be estimated to be within the range of 135 to 180 minutes (about 3 hours).

In comparison to the manual method, our model's web scraping process for collecting each faculty's information and saving it locally takes less than a minute. The extraction of information, which varies in time depending on the process capability of each computer, takes an additional minute.

Subsequently, presenting the extracted information in the form of knowledge graphs takes about another minute. Notably, our model can operate concurrently and collect data on multiple faculty members simultaneously. As such, the entire process of collecting and saving information from all nine faculty members at CGU CISAT using our model takes less than three minutes. In comparison to manual methods, our model offers significant time-saving advantages. In the case of our example, we also can expand the number of instances from a small sample to a larger amount to further evaluate the effectiveness of the model. For instance, while our study was limited to examining nine faculties at CGU Information Technology department, we can apply the same methodology to a much larger public school such as UCLA (University of California, Los Angeles), which has over 50 faculties. By doing so, we can quantify how much time the model can save when scraping 50 faculties compared to collecting their profiles manually.

In terms of accuracy, our model's evaluation is based on its ability to extract and present relevant entities and relationships from the data. As such, our model accuracy rates rely on the accuracy of the Information Extraction process. Inaccurate information extraction can lead to incorrect or incomplete results, which can significantly affect the system's overall performance. Therefore, ensuring high accuracy in information extraction is crucial for our model achieving reliable and effective results. To evaluate information extraction accuracy, several measures are commonly used, such as precision, recall, and F1-score. Precision refers to the percentage of correctly extracted information compared to the total number of extracted information, while recall refers to the percentage of correctly extracted information compared to the total number of relevant information in the input data. F1-score is a combination of precision and recall, which provides an overall evaluation of the system's performance. As an example, when extracting information for Dr. Chipidza, our model was able to identify and extract 18 entities and 14 relationships. Note that the accuracy of relationship extraction is not a fixed result and can be subject to change. This is influenced by factors such as the quality of the training datasets and the effectiveness of the information extraction models used. In our case, we employed the REBEL model for the Information Extraction process. It is important to note that the IE results can be affected by the training datasets used for the REBEL model, and the results can also differ when using other IE models. This variability can result in fluctuating accuracy rates for the model. Fortunately, there is one thing that we can be certain about. With the continuous development and advancements in artificial intelligence technology, we expect to see a constant improvement in the accuracy of information extraction capabilities of our model in future.

6.3 Social-technical Evaluation

The social-technical evaluation process in DSR is typically divided into two parts: quantitative evaluation and qualitative evaluation (Hevner & Chatterjee, 2010). Quantitative evaluation, such as survey, involves collecting numerical data to measure the effectiveness of the artifact. Qualitative evaluation involves collecting non-numerical data such as interviews to understand the user experience and to gain insights into how the artifact can be improved. Both quantitative and qualitative evaluation methods are essential in social evaluation. Quantitative evaluation provides objective data that can be analyzed to determine the artifact's effectiveness, while qualitative evaluation provides valuable insights into the user experience and how it can be improved. The combination of these two methods ensures that the artifact is thoroughly evaluated and refined to address the intended problem.

There are various methods for conducting social evaluations, including surveys, interviews, focus groups, and others. We will use both quantitative surveys and qualitative interviews for social-technical evaluation to examine our model in this study.

Quantitative Surveys

In this study, we employed a questionnaire survey to evaluate the usefulness of our model. The survey was conducted with 10 volunteers asked to provide feedback on our model. The aim was to investigate the impact of information overload in modern times and assess whether our model was effective in addressing the problem.

The survey revealed that most of the participants (90%) believed we are currently experiencing an age of information explosion and that information overload has become a significant challenge in modern times. Over 50% of the respondents also reported encountering information overload frequently, with only 10% stating that they do not experience information overload problems. Respondents reported limited methods to manage information overload, including limited technologies, meditations, and others. Additionally, we presented our model and two use cases to the survey participants, and over 60% of the respondents reported that the information flow model was very useful, with 10% stating that the model was extremely useful.

Qualitative Interviews

To gain a more comprehensive understanding of our model and obtain detailed feedback, we conducted one focus group interview and two in-depth interviews.

Focus group interviews are a qualitative research technique that involves a small group of people who are brought together to discuss a particular topic or issue. The purpose of a focus group interview is to gain insight into the perceptions, opinions, and attitudes of the participants towards the topic being discussed. For our research, we conducted a qualitative interview with a small focus group (three people). Although the group had not been previously exposed to the term "information overload," they expressed their belief that they had experienced many situations where they were overwhelmed with information in their daily lives. For instance, one participant, who was a parent, shared his experience of “I have to read over 50 pages of pdf documents, the first day for my daughter’s kindergarten”. During our discussion about how they manage information overload and what technology they use for information management, we found that the only tool being used by the participants was Google search. The same parent stated that “50 pages of pdf, I simply just discard or ignore the information, or ask the other parents who ever read the pdf documents”. Upon presentation of the model to the group, they expressed their appreciation for the concept of the model and believed it would be highly beneficial if the model were fully intelligent. Regarding information visualization, the participants felt that the knowledge graph was too complex due to the numerous relationships between different entities. This led to a question about what is the balance between model accuracy and model complexity? While high accuracy is desirable, extracting too many entities and relationships may not be the goal for our model. It may be necessary to focus on presenting only the most important entities and relationships to avoid overcrowding the output graphs with noise information and keep them simple for our future research.

In this research, we also conducted two in-depth interviews with two IT experts, in which each interview lasted about 1.5 hrs. to 2 hrs. An in-depth qualitative interview is a research method used to gather detailed and comprehensive information about a specific topic or phenomenon. It involves conducting a one-on-one interview with a participant or a group of participants in a structured or semi-structured manner. The purpose of an in-depth qualitative interview is to gain a deep understanding of the participant's experiences, beliefs, attitudes, and perspectives related to the topic under investigation. The participants provided rich and detailed insights into the experiences and perspectives of participants, which can be used to inform further research or practical applications.

One of the participants, Manar, a PhD researcher in Information Technology from CGU, strongly desired to share her story and insights about information overload by her real name. Manar defines information overload as the point where the amount of information about a specific topic, matter, or field exceeds the human capacity to manage it. She emphasizes that information overload is a significant issue that can result in overthinking, and in some cases, can cause enormous stress in daily life, leading to insomnia and even mental health issues. Then Manar shared her personal experience about information overload that she had suffered from information overload in the last three years and believes that information overload has become the primary obstacle hindering her from achieving her goals, success, and dreams. Upon discussing the methods, she uses to manage information overload, it was revealed that she only relies on basic software (such as Microsoft Excel, etc.) and Internet search engines, as there are no other high-tech options available. She mentioned that for the past three years, she has been searching for software or consultants to aid her in managing information overload. Upon presentation of the model, she expressed her belief that it would be highly beneficial for information management. Also, as part of her feedback on improving the model, she suggested incorporating an information summary phase to the model. As she believes such a phase would benefit users by reducing their reading time. Furthermore, she emphasized two critical points that the model should pay attention to: firstly, ensuring that the model does not display any bias and accurately represents the facts during information extraction. Secondly, it must provide precise translations without any misleading information.

The other participant, Felipe, an IT manager with 15 years of experience at Disney and a PhD in Information Systems from CGU, is extremely impressed by the model and wishes to contribute his knowledge under his real name. Felipe also believes that information overload is a significant challenge as it can make it difficult to derive any meaningful insights from the flood of data.

Then Felipe shared his personal experience about information overload and gave an example from his work. When he worked as a consultant, he was faced with an overwhelming amount of information during his initial interaction with a client in a short span of time. Felipe also emphasized that such an information overload can have a profound impact on one's life causing them to constantly think about the information even while sleeping, and affecting their daily activities, such as socializing and eating. Felipe's strategy for managing information overload involves breaking down tasks into manageable portions, but he faces difficulties in finding the appropriate IT tools to support him. Currently, he relies on limited IT techniques such as Evernote to help him manage information overload.

Upon being introduced to our information process model, Felipe was extremely impressed by the model and expressed his appreciation for the model's usefulness, timesaving potential, and overall interest. Particularly, he highlighted the knowledge graph network component, which he described as great and fascinating. Felipe believes that knowledge graphs can extract more valuable information, particularly in terms of relationships between entities, and that they are more efficient than traditional reading methods. According to him, when humans read, they often have to read multiple times to understand the relationships, which is unnecessary with knowledge graphs. He further noted that the graphs are like maps, providing guidance and helping to visualize the big picture of the information, meanwhile providing many detailed level information.

Furthermore, Felipe expressed his belief that our information process model has a wide range of potential applications beyond the current context, and he sees it as having significant potential for commercial use. He cited CRM (Customer Relationship Management) as an example, stating that the model would be highly beneficial for managing customer networks. This approach can be used to better understand customer needs, customer preferences, etc. and provide valuable insights for businesses. Overall, the potential for commercial use is vast, and the model could be applied in numerous areas.

In conclusion, the evaluation of model performance is crucial in determining its effectiveness in real-world scenarios. Our study used both technical evaluation and social evaluation (quantitative and qualitative) processes that provide a comprehensive assessment of the model's performance. The results demonstrate that the model is efficient in terms of time efficiency and effective in addressing the challenge of information overload. By using our model, users can collect valuable information as well as extract high-quality information more effectively and efficiently in the information explosion age. Furthermore, the model has great potential for commercial use.

Therefore, the model can be considered successful in addressing the challenge of information overload. In the following section, we will discuss the potential prospects of our model for future research.

Chapter 7 Discussions

Through the utilization of Python libraries, we have successfully implemented the use cases into our model. The incorporation of Python libraries has proven to be instrumental in achieving the desired outcome, allowing us to effectively address the research question at hand. The successful integration of these libraries has enabled us to analyze and process the textual data to KG data model efficiently, leading to meaningful and valuable results.

By using our model, users can collect information more efficiently, and the model also enables visualization of textual information by transforming unstructured textual data into a structured knowledge-based data model. The successful integration of those Python libraries into the model builds an automatic information flow that has furthered our understanding of the research problem and has also opened new possibilities for future research in this area.

7.1 Potential Further Work

Using Knowledge Graphs to present key information constitutes a notable advantage of our proposed model. By leveraging machine learning and natural language processing (NLP) techniques, the Knowledge Graph extracts entities from unstructured data and establishes relationships between them within a schema that forms the foundation of the graph. The resulting data model captures the interconnectedness of entities and their relationships, providing a simplified and visual representation of complex layers of knowledge. It is important to note that a Knowledge Graph surpasses the conventional notion of a knowledge base, as it embodies a graphical depiction of an organization's knowledge base in the form of a digital network of interconnected data entities.

The potential of our model is significant, as simple graphs can be extended to form a large network with many Knowledge Graphs. This implies that the scalability and expansiveness of our model is valuable, allowing for the representation of vast and complex knowledge domains. The ability to represent diverse entities and their relationships within a graphical framework empowers our model to effectively capture and convey knowledge in a comprehensive and accessible manner. This underscores the promising prospects of our proposed model in facilitating efficient knowledge representation and retrieval, and has the potential to contribute to advancements in various domains where information organization and visualization are critical, for example, Wikipedia Knowledge Base, Google search, etc.

From Wikipedia to Knowledge Graphs Network

A Transition from Wikipedia to Knowledge Graphs Network: An Entity Network Perspective.

Wikipedia is a free online encyclopedia that allows users to create and edit articles collaboratively. It has become one of the most widely used sources of information on the Internet, with millions of articles on assorted topics. In Wikipedia, information is presented as articles that may contain unstructured information. It is up to the reader to identify the key entities and their relationships within the article. This can be a time-consuming and error-prone process, especially for complex topics that require the integration of multiple articles.

The transition from Wikipedia to Knowledge Graphs Network represents a significant shift in the paradigm of knowledge representation. In the traditional Wikipedia model, knowledge is organized in the form of articles, whereas in a Knowledge Graphs Network, entities and their relationships are organized in a graphical network. This transition involves the adoption of advanced techniques in data extraction, entity recognition, and relationship mapping, which enable the representation of knowledge in a more interconnected and contextual manner. The Entity Network perspective emphasizes the evolution from a static Wikipedia-based knowledge representation to a dynamic and interconnected Knowledge Graphs Network, which has the potential to revolutionize how information is organized and retrieved in various domains.

From Google Search to Knowledge Graphs Search

Evolving from Google Search to Knowledge Graphs Search: Emphasizing Entity and Relationship Queries.

Google Search has historically been the dominant search engine used by individuals seeking information online. It has relied on a keyword-based search methodology, where users input keywords or phrases, and the search engine retrieves relevant results by matching these keywords with the content of web pages. Nevertheless, this conventional approach has inherent limitations in comprehending the context, meaning, and interrelationships among entities, which can lead to imprecise or incomplete search results. The restricted focus on keywords fails to capture the nuances of language, semantics, and the complex relationships between entities, resulting in a partial representation of the knowledge available on the web.

The evolution from Google Search to Knowledge Graphs Search signifies a paradigm shift in information retrieval approaches. In traditional Google Search, queries are primarily based on keywords and search results are ranked based on relevance. However, with the advent of Knowledge Graphs, the emphasis has shifted towards entity and relationship-based queries, where entities are recognized, and their relationships are mapped in a graphical network. This shift involves the utilization of advanced techniques in natural language processing, machine learning, and graph-based algorithms, which enable a more context-aware and semantic-driven search experience. The transition towards Knowledge Graph Search underscores the increasing importance of entities and their relationships in information retrieval and has the potential to revolutionize how users interact with search engines and access relevant information.

7.2 Model Limitations

During our case study evaluation section, we discussed the significant challenges when attempting to apply to our information-processing model. These challenges included difficulties on web scraping during the information collection process, challenges arising from the information extraction process, and challenges related to information visualization. Additionally, there were challenges associated with building an automatic information pipeline that could allow for seamless information flow throughout the entire model. Despite the challenges we had, the model also has some limitations.

Our model provides a systematic approach to developing and implementing IT artifacts for automatic data flow; however, a static model may become outdated as time progresses.

The primary concern of our model is that it is not time and environment sensed. This means that the model may become outdated over time, resulting in reduced effectiveness and reliability. A static model may not be able to adapt to modern technologies or changing requirements, which can result in a loss of competitive advantage. To maintain the effectiveness and relevance of the model, the IT artifacts need to be updated or upgraded as time progresses. This means that the model artifacts and innovation need to evolve with time.

A dynamic model is a potential solution for the challenges associated with the current static framework. The dynamic model will be time and environment-sensing that can adapt to changes in the environment and real time requirements. This will make the model more flexible and responsive to changes, ensuring that it remains relevant and effective over time.

7.3 Contributions

The success of integrate IT artifacts, information collection, information extraction, and information visualization to an automated information process pipeline model has contributed an innovative knowledge to both Design Science Research and Information System and Technology field.

The successful implementation of case studies to our model not only greatly enhanced the validity and rigidity of our study but also presented a practical real-world demonstration for the DSR community, IST community, and Society.

7.3.1 Knowledge Contributions

DSR Theory and DSR Community

Design Science Research (DSR) is an approach to conducting research that aims to develop and design IT artifacts that can solve practical problems. DSR focuses on creating and testing new models, frameworks, and systems that can contribute to the advancement of the IT field. The paper has explored the concept of artifacts in Design Science Research (DSR), in which, artifacts refer to any material or artificially made object that can be transformed into a tangible existence, such as models, instantiations, or processes like software and methods (Goldkuhl, 2002).

There are two types of knowledge bases in DSR, which are descriptive and prescriptive knowledge (Gregor & Hevner, 2013). Descriptive knowledge describes natural, artificial, and human phenomena, including observation, classification, measurement, and cataloging. Prescriptive knowledge, on the other hand, offers solutions or prescriptions for problems through constructs, models, methods, instantiations, and design theory. We developed an information process model to resolve information overload problems, which align with prescriptive knowledge. There are four outcomes of DSR, which are constructs, models and frameworks, methods and algorithms, and instantiations (March & Smith, 1995). Our delivery falls into the category of models and frameworks. There are three types of DSR contribution levels: situated implementation of artifacts, nascent design theory, and well-developed design theory. Situated implementation of artifacts refers to software products or implemented processes, while nascent design theory pertains to knowledge as operational principles/architecture, including constructs, methods, models, design principles, and technological rules. Lastly, well-developed design theory involves design theories about embedded phenomena (Gregor & Hevner, 2013). Our model belongs to level 2 nascent design theory.

Thus, in this study, we presented an automated information process model and developed related use cases to address the problem of information overload. The model artifacts belong to prescriptive knowledge, which means that it provides a set of guidelines or recommendations for tackling the issue. The theory contribution type falls under level 2 Nascent Design Theory, which means that the researcher has developed a new model that has not yet been fully tested but has significant potential for practical applications. The automated information process pipeline model has significantly contributed to the DSR community as it presented a successful DSR case. The model has demonstrated its feasibility and effectiveness in tackling information overload, and it has the potential to inspire and guide DSR researchers to design and create more efficient and practical systems that can be applied in real-world scenarios. Our research contribution to the DSR community is significant as we have developed a new model that can help solve practical problems related to information overload. The model provides a prescriptive approach that can be applied in different settings and industries, and it can be extended or customized to address other related issues. Moreover, the successful implementation of the model in real-world use cases can inspire further research and development in the DSR field. DSR researchers can learn from the experiences of the researchers in this study and use the automated information process model as a basis for developing new models or frameworks that can contribute to the advancement of the IT field. Thus, the research contribution to the DSR community is significant, as we have presented an innovative and practical solution to the problem of information overload. The automated information process model can inspire further research and development in the IT field and guide researchers to design and create more efficient and practical systems that can be applied in real-world scenarios. The researchers have successfully demonstrated the feasibility and effectiveness of the model in real-world use cases, making a significant contribution to the DSR community.

IST Field and IST Community

The field of Information Systems and Technology (IST) has been plagued with the issue of information overload for many years. Researchers have been trying to design and develop effective tools and techniques to help manage this problem. In this context, the current study presents an automated information process model to tackle the issue of information overload, which has made significant contributions to the IST field. The automated information process model also has contributed innovative knowledge to the IST field as it provides a practical and efficient way to convert unstructured textual information into a knowledge graph data model. At the same time, it presented a comprehensive set of tools and techniques that can facilitate data collection, analysis, and visualization processes. Most importantly, it provides valuable guidance or roadmap for researchers and practitioners in the IT field for the future. This model can be extended or applied to other IT artifacts, for example, they can extend the model to Google search or build a Wikipedia Entity Network. The model has significant contributions to the IST community. Our model will guide and inspire more IT researchers to do further investigations on the Information Extraction and Data Visualization process. We believe more innovative systems related to knowledge graphs will be developed in the future. Much more advanced IT artifacts may be coming out in the future after a successful real world case study implementation of our model.

Therefore, the current study has made significant contributions to the IST field. The automated information process model presented in the study has the potential to help manage information overload issues faced by many individuals and organizations. This model's contributions are valuable, and it provides guidance for future research in the IT field. Further investigations on the

Information Extraction and Data Visualization process may lead to more advanced IT artifacts that can tackle the issue of information overload more effectively.

7.3.2 Real World Contribution

The use case studies conducted to evaluate the model's feasibility and validity showed its effectiveness and real-world applicability. Our model has the potential to be used in a variety of contexts, both within society and in commercial organizations. Its ability to efficiently manage and extract valuable information is an asset that could benefit a wide range of users. Based on our survey and qualitative interviews, it is apparent that many individuals and organizations have experienced some degree of information overload. Despite this, there are limited IT software solutions available to help them manage this issue. Therefore, our model presents an opportunity for high technology solutions to be developed and implemented to help individuals and organizations manage their information overload issues. Furthermore, the potential for our model to be developed for commercial use is an exciting prospect. It has the potential to help any users or organizations that are suffering from information overload, making it an asset for a wide range of industries. The benefits of this could be significant, ranging from increased efficiency to improved decision-making capabilities.

In conclusion, our information process model can make a real-world contribution to addressing information overload. Its effectiveness and real-world applicability have been demonstrated through use case studies, and its potential for commercial use presents an exciting opportunity for further development and implementation. By providing a solution to information overload, our model could have a significant positive impact on society and commercial organizations alike.

Chapter 8. Conclusions and Future Work

8.1 Conclusions

To conclude, this paper has identified the problem of information overload in the digital age, which has created significant challenges for researchers who struggle to find and extract useful information from enormous amounts of unstructured text data.

We explored existing solutions, which rely on personal commitment and contextual strategies that may not be generalizable to all conditions, and we also found the existing solutions lack details on how IT techniques can help to alleviate the problem of information overload.

To address these limitations, we proposed an automated information process model based on design science theory. Our approach involves utilizing an automated information and data flow framework that incorporates modern information techniques such as web scraping, natural language processing, and knowledge graphs. By doing so, we aimed to streamline the collection of essential information while also optimizing the conversion of unstructured text data to a structured data model. Then we applied three real use cases to our model and conducted both quantitative and qualitative evaluation processes to assess its effectiveness. The results indicated that our model is practical and can significantly reduce the burden of information overload for researchers.

We also discussed possible potential future work for model extension, transition from Wikipedia to Knowledge Graphs Network and evolving from Google Search to Knowledge Graphs Search.

Overall, this paper contributes to the ongoing efforts to tackle the problem of information overload, and we hope that our proposed model will inspire further research and innovation in this area.

8.2 Future Work

In this study, we presented an automated information process model that can help tackle information overload, and then developed related use cases to demonstrate the model's effectiveness. However, there are still some areas where further research and development are needed.

One of the futures works we plan to do is to develop a dynamic model that can keep up with the rapid development of new IT techniques. The IT industry is constantly evolving, and innovative technologies are emerging all the time. Therefore, it is crucial for the model to stay up to date to remain effective. By developing a dynamic model, the researchers can ensure that the model remains relevant and useful for tackling information overload in the future.

Another area where we plan to focus on in the future is developing more case studies and demonstrations to evaluate the validity, feasibility, and practicability of the framework. Case studies and demonstrations can provide valuable insights into how the model performs in different real-world scenarios. By conducting more case studies and demonstrations, the researchers can gather more data to validate and refine the model further.

Furthermore, we plan to extend the model to Wikipedia and Google. We will conduct more research on transitioning Wikipedia to Knowledge Graphs Network and evolving Google Search to Knowledge Graphs Search. This extension of the model will enable us to explore how the model can be applied to other IT artifacts and further expand its applicability.

In conclusion, the future work for this study involves developing a dynamic model, conducting more case studies and demonstrations, and extending the model to other IT artifacts. By doing so, we can continue to refine and improve the model, making it more effective in tackling information overload and contributing to the advancement of the field of information systems and technology.

Bibliography

Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, *11*. https://doi.org/10.1177/1847979019890771

Agrawal, R. (2021, October 8). Beginner's guide on how to prepare dataset using web scraping. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/10/beginners-guide-on-how-to-prepare-dataset-using-web-scraping/

Babelscape. (2021). *REBEL*. GitHub. https://github.com/Babelscape

Belfourd, T., & Furner, J. (1997). Fast learners or time wasters? Intelligent agents on the Web: A user study. *Managing Information*, *4*(9), 32–34.

Bielska, A. (2020). *Open Source Intelligence Tools And Resources Handbook 2020*. https://i-intelligence.eu/uploads/public-documents/OSINT_Handbook_2020.pdf

Borhade, V. (2021, April 19). Knowledge Graph—Future of Data. *Medium*. https://varshaborhade.medium.com/knowledge-graph-future-of-data-464b933a4228

Brennan, L. L. (2011). The scientific management of information overload. *Journal of Business and Management*, *17*(1), 121–134.

Brian Solis. (2020, June 17). *Information overload, why it matters and how to combat it*. The Interaction Design Foundation. https://www.interaction-design.org/literature/article/information-overload-why-it-matters-and-how-to-combat-it

Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2256–2262. https://doi.org/10.18653/v1/D16-1245

Darling, G. (2023, February 10). *Is chatgpt reliable? Here's why it's not • thecodebytes*. Thecodebytes.

https://thecodebytes.com/is-chatgpt-reliable-heres-why-its-not/

David, B. (2022). *Crosslingual Coreference—A Multi-Lingual Approach to AllenNLP CoReference Resolution along with a Wrapper for SpaCy*. https://github.com/Pandora-Intelligence/crosslingual-coreference.

Djuraskovic, O. (2021). 30+ Big Data Statistics (2023)—Amount of Data Generated in The World. *FirstSiteGuide*. https://firstsiteguide.com/big-data-stats/.

Domino Data Science Dictionary. (2023). *What is spacy?* https://www.dominodatalab.com/data-science-dictionary/spacy

Dunnigan, J. (2002). Open source intelligence. In M. M. Lowenthal (Ed.), *Intelligence: From Secrets to Policy*. CQ Press.

Edmunds, A., & Morris, A. (2020). The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, *20*(ue 1). https://doi.org/10.1016/S0268-4012(99)00051-1

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The Information Society*, *20*(5), 325–344. https://doi.org/10.1080/01972240490507974

ExpressAnalytics. (2022, July 1). *Named Entity Recognition (NER): Benefits, Use Cases, Algorithms*. https://www.expressanalytics.com/blog/what-is-named-entity-recognition-ner-benefits-use-cases-algorithms/

Fulton, C., & McGuinness, C. (2016). *Digital detectives: Solving information dilemmas in an online world*. Chandos Publishing.

Goldkuhl, G. (2002). Anchoring Scientific Abstractions—Ontological and Linguistic Determination Following Socio- Instrumental Pragmatism. *European Conference on Research*.

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum

Impact". *MIS Quarterly*, *37*(2), 337-55,. https://doi.org/10.25300/misq/2013/37.2.01.

Hassan, N. A., & Hijazi, R. (2018). Open Source Intelligence Methods and Tools. *Apress*. https://douran.academy/wp-content/uploads/ebooks/open-source-intelligence-methods-and-tools.pdf

Hemp, P. (2009, September 1). Death by information overload. *Harvard Business Review*. https://hbr.org/2009/09/death-by-information-overload

Hevner, A. R., & Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice*. Springer.

Hevner, March, Park, & Ram. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625

Huggingface. (2023). *NeuralCoref 4.0: Coreference Resolution in spaCy with Neural Networks.* [C]. Hugging Face. https://github.com/huggingface/neuralcoref (Original work published 2017)

Huguet Cabot, P.-L., & Navigli, R. (2021). Rebel: Relation extraction by end-to-end language generation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2370–2381. https://doi.org/10.18653/v1/2021.findings-emnlp.204

Igraph. (R Interface). https://r.igraph.org/articles/igraph.html.

Introducing chatgpt. (2023). https://openai.com/blog/chatgpt

IORG. (2022). *About iorg | information overload research group*. IORG. https://iorgforum.org/about-iorg/

Jain, R. (2018). Natural Language Extraction- Using SpaCy on a Set of Novels. *Medium*. https://medium.com/@rajat.jain1/natural-language-extraction-using-spacy-on-a-set-of-novels-88b159d68686.

Ji, H. (2009). Information Extraction. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems*. Springer. https://doi.org/10.1007/978-0-387-39940-9_204

Huettich, J. (2020, February 10). The ultimate guide to managing information overload. *MindManager Blog*.

https://blog.mindmanager.com/202002the-ultimate-guide-to-managing-information-overload/

Jiménez, H. (2022). From text to a knowledge graph. *Hands-on. Medium*. https://medium.com/@hugojm/from-text-to-a-knowledge-graph-hands-on-dd68e9d42939

JNordine. (2022). *Osint framework*. https://osintframework.com/

Joshi, P. (2019, September 23). Information extraction | information extraction python,spacy. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2019/09/introduction-information-extraction-python-spacy/

Klassen, T. P., Jadad, A. R., & Modher, D. (1998). Guides for reading and interpreting systematic reviews. *Archives of Pediatrics and Adolescent Medicine*, *152*(7), 700–704.

Kurama, V. (2021). What Is Information Extraction? - A Detailed Guide. *Nanonets AI & Machine Learning Blog*. https://nanonets.com/blog/information-extraction/.

Laskin, D. M. (1994). Dealing with information overload. *Journal of Oral Maxillofacial Surgery*, *54*(7), 661.

Majumder, P. (2021, June 16). Named entity recognition (NER) in python with spacy. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

McDowell, D. (2008). *Strategic Intelligence*. Scarecrow Press.

Meyer, P. (2023, January 27). *Automatic knowledge graphs: The impossible grail*. Medium. https://pub.towardsai.net/automatic-knowledge-graphs-the-impossible-grail-ef71f9c8aad8

Minner, S. D. (2018). Overcoming Information Overload: Open Source Intelligence in a Modern Threat Environment. *MARINE CORPS UNIV QUANTICO VA*.

Named entity recognition: Concept, tools and tutorial. (2020, March 30). MonkeyLearn Blog. https://monkeylearn.com/blog/named-entity-recognition/

Nayak, T., & Ng, H. T. (2020). Effective modeling of encoder-decoder architecture for joint entity and relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8535.

NetworkX. (2023). *NetworkX — NetworkX documentation*. https://networkx.org/

Nltk: Natural language toolkit. (2023). https://www.nltk.org/

Norton, Miller, M., & Meyer, P. (2015). *Heterogeneous Graph Recommendation Model based on Graph Neural Network*. https://doi.org/10.21203/rs.3.rs-2102785/v1.

Ontotext. (2023). What is a Knowledge Graph? *Ontotext*. https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

Patel, H. (2022, October 21). *How web scraping is transforming the world with its applications*. Medium. https://towardsdatascience.com/https-medium-com-hiren787-patel-web-scraping-applications-a6f370d316f4

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Persson, E. (2019). *Evaluating tools and techniques for web scraping*. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-271206

R., R. (2017). *The Effect of Information Overload in Digital Media News Content*.

Sajid, H. (2022, November 1). *8 benefits of building a knowledge graph | wisecube ai*. https://www.wisecube.ai/blog/8-benefits-of-building-a-knowledge-graph/

Ruder, S. (2023). *Relationship extraction*. NLP-Progress. http://nlpprogress.com/english/relationship_extraction.html

Shankar. (2022, May 31). A complete guide on feature extraction techniques. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/05/a-complete-guide-on-feature-extraction-techniques/

Simon, H. A. (1996). *The Sciences of the Artificial* (3rd ed.). MIT Press.

Singh, S. (2018). *Natural language processing for information extraction*.

Singh, S. (2019). Analyzing coreference tools for NLP application. *International Journal of Computer Sciences and Engineering*, *7*(5), 608–615.

Spacy Universe. (2023). *Linguistic features · spacy usage documentation*. Linguistic Features. https://spacy.io/usage/linguistic-features

Spacy Universe. (2023b). *Spacy 101: Everything you need to know · spacy usage documentation*. SpaCy 101: Everything You Need to Know. https://spacy.io/usage/spacy-101#whats-spacy

Spacy Universe: Crosslingual coreference. (2023). *Crosslingual coreference · spacy universe*. Crosslingual Coreference. https://spacy.io/universe/project/crosslingualcoreference

Stanford edu. (2023). *What is a knowledge graph?* Stanford edu. https://web.stanford.edu/~vinayc/kg/notes/What_is_a_Knowledge_Graph

Singh, T. (2023, January 2). *Natural language processing with spacy in python – real python*. https://realpython.com/natural-language-processing-spacy-python/

Taylor, P. (2022, September 8). *Total data volume worldwide 2010-2025*. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

The stanford natural language processing group. (2023). https://nlp.stanford.edu/software/CRF-NER.shtml

Wikipedia. (2022). Open-source intelligence. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open-source_intelligence&oldid=1116471597

Wikipedia. (2023). Information extraction. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Information_extraction&oldid=1135544307

Wilson, P. (1997). Interdisciplinary research and information overload. *Library Trends*, *45*(2), 192–203.

Zenrows. (2023). *Selenium vs beautifulsoup in 2023: Which is better? - zenrows*.

https://www.zenrows.com/blog/selenium-vs-beautifulsoup

Appendix

1. Open-Source Intelligence Search Tools

- Google AdWords
- Keyword Tool
- KWFinder
- Keyword discover
- Keyword Shitter
- One Look: Enter a word, phrase, sentence, or pattern to search for related words.
- Ubersuggest: suggest keywords not available in the Google Keyword Planner.

Files Search Engines

- Fagan Finder
- DOCUMENT SEARCH ENGINE
- greyhat warfare: Search for Open Amazon s3 Buckets and their contents.

2. Web Scraping Library

Python:

Scrapy: Scrapy is an open-source web scraping framework written in Python. It provides a set of tools and libraries that allow you to easily and efficiently extract data from websites. Scrapy handles the complexities of crawling, parsing, and storing data, making it a popular choice for web scraping tasks. Key features of Scrapy:

- Fast and efficient: Scrapy is built on top of the Twisted asynchronous networking framework, which allows it to handle large amounts of concurrent requests efficiently.

- Built-in support for handling common web scraping tasks: Scrapy provides functionality for handling pagination, following links, submitting forms, handling cookies, and managing user sessions.
- Extensibility and modularity: Scrapy's architecture is designed to be highly modular and extensible. You can customize and extend Scrappy's functionality by implementing your own middlewares, pipelines, and extensions.
- Item Pipeline: Scrapy has a built-in item pipeline that allows you to process scraped items and apply different transformations or validations before storing them.
- Scrapy Shell: Scrapy provides an interactive shell, known as the Scrapy Shell, which allows you to quickly test and debug your scraping code.

BeautifulSoup: Beautiful Soup is a Python library used for web scraping and parsing HTML or XML documents. It provides a convenient and flexible way to extract data from web pages by navigating and searching through the HTML structure. With Beautiful Soup, you can parse HTML or XML documents and extract specific elements, such as tags, attributes, and their content. It allows you to navigate the parsed document using methods and properties that make it easy to search for specific elements or access their data. Key features of Beautiful:

- Parsing: Beautiful Soup can handle poorly formatted or broken HTML, making it resilient to imperfect web page structures.
- Tag and attribute searching: You can search for specific HTML tags or attributes within the parsed document, allowing you to target specific elements.
- Data extraction: Beautiful Soup provides methods to extract the text content, attributes, or other data associated with HTML elements.

- Traversing the HTML tree: You can navigate the HTML structure by accessing parent, sibling, or child elements, enabling you to explore and extract data hierarchically.
- Integration with parsers: Beautiful Soup supports various parsing libraries, including Python's built-in "html.parser" and external libraries like lxml and html5lib.

MechanicalSoup: MechanicalSoup is a Python library that simplifies the process of automating interactions with websites. It acts as a higher-level wrapper around the popular requests and BeautifulSoup libraries, providing an intuitive API for web scraping and form submission tasks. MechanicalSoup combines the capabilities of requests, which is used for making HTTP requests, and BeautifulSoup, which is used for parsing HTML or XML documents. It allows you to automate the process of filling out and submitting forms on websites, as well as extracting data from the resulting pages. Key features of MechanicalSoup:

- Form submission: MechanicalSoup enables you to easily fill out and submit HTML forms on websites. It handles the underlying HTTP requests and cookies required for form submission.
- Browser-like interaction: MechanicalSoup emulates the behavior of a web browser, allowing you to navigate through pages, follow links, and interact with forms.
- Integration with BeautifulSoup: MechanicalSoup seamlessly integrates with BeautifulSoup, enabling you to parse and extract data from the HTML or XML content of web pages.
- Session management: MechanicalSoup provides session management capabilities, allowing you to persist cookies and maintain state across multiple requests.
- Easy-to-use API: MechanicalSoup aims to provide a simple and intuitive API for common web scraping tasks, making it easier for beginners to get started.

R:

Rvest: Rvest is an R package that facilitates web scraping and extracting data from websites in the R programming language. It provides a simple and intuitive API for navigating and parsing HTML or XML documents, making it easier to scrape and extract information from web pages. key features of Rvest:

- HTML/XML parsing: Rvest utilizes the xml2 package to parse HTML or XML documents, allowing you to extract data from the parsed structure.
- Selectors: Rvest supports CSS selectors, allowing you to target specific HTML elements based on their tags, classes, or attributes.
- Data extraction: Rvest provides functions to extract data from HTML elements, such as text content, attributes, or specific element properties.
- Navigation: Rvest allows you to navigate through the HTML structure, accessing parent, sibling, or child elements to explore and extract data hierarchically.
- Form submission: Rvest enables you to fill out and submit HTML forms on websites, automating the process of interacting with web forms.
- Session management: Rvest provides session management capabilities, allowing you to maintain cookies and state across multiple requests.

Java:

Jaunt: Jaunt is a Java library that simplifies web scraping and automation tasks in the Java programming language. It provides a high-level API for interacting with websites, making it easier to extract data, navigate through pages, and automate web interactions. Key features of Jaunt:

- HTML parsing: Jaunt supports parsing and extracting data from HTML or XML documents. It provides methods to access specific elements, attributes, and text content.

- Form submission: Jaunt allows you to fill out and submit HTML forms on websites. It handles the underlying HTTP requests and manages cookies and sessions.
- Session management: Jaunt provides session management capabilities, allowing you to persist cookies and maintain state across multiple requests.
- Navigation: Jaunt allows you to navigate through the HTML structure, accessing parent, sibling, or child elements to explore and extract data hierarchically.
- JavaScript execution: Jaunt has built-in support for executing JavaScript code within web pages, enabling you to interact with dynamically generated content.
- Proxy support: Jaunt provides functionality to work with proxies, allowing you to route requests through different IP addresses for anonymity or geolocation purposes.
- Easy-to-use API: Jaunt aims to provide a simple and intuitive API for web scraping and automation tasks in Java, making it easier for developers to get started.

JSoup: Jsoup is a Java library that facilitates HTML parsing, manipulation, and extraction of data from web pages. It provides a convenient and flexible API for working with HTML documents, allowing developers to easily navigate, modify, and extract information from HTML content. Key features of Jsoup include:

- HTML parsing: Jsoup parses HTML documents, handling imperfect or poorly formatted HTML gracefully. It can parse HTML from files, URLs, or plain text.
- DOM traversal: Jsoup allows developers to traverse the HTML document using methods and selectors similar to those in CSS or jQuery. This makes it easy to locate specific elements or extract data from the document.
- Element manipulation: Jsoup enables developers to modify HTML elements, add or remove attributes, change text content, or manipulate the HTML structure as needed.

- Data extraction: Jsoup provides methods to extract specific data from HTML elements, such as retrieving text content, attribute values, or element properties.
- HTML sanitization: Jsoup includes built-in functionality to sanitize HTML input by removing potentially harmful content, helping prevent cross-site scripting (XSS) attacks.
- Form submission: Jsoup supports filling out HTML forms and submitting them to websites. It can handle cookies, manage sessions, and handle redirects during form submission.
- Integration with CSS-like selectors: Jsoup allows developers to use CSS-like selectors to target specific HTML elements based on their tags, classes, or attributes.

HtmlUnit: is a headless Java browser allowing commonly used browser functionality such as following links, filling out forms and more. HtmlUnit is a Java library that acts as a headless browser for simulating web browser behavior programmatically. It allows developers to interact with web pages, execute JavaScript, and extract data from websites without the need for a graphical user interface (GUI). Key features of HtmlUnit include:

- Headless browsing: HtmlUnit provides a headless browser environment, which means it operates without a visible browser window. This makes it suitable for server-side or automated web scraping tasks.
- Web page navigation: HtmlUnit supports navigating through web pages, following links, and interacting with forms, just like a regular web browser.
- JavaScript execution: HtmlUnit has built-in JavaScript support, allowing it to execute JavaScript code embedded within web pages. This enables developers to interact with dynamic and AJAX-driven websites.

- HTML parsing: HtmlUnit parses HTML documents, allowing developers to extract data from the parsed structure. It provides methods to access specific elements, attributes, and text content.
- Cookie and session management: HtmlUnit supports handling cookies and maintaining session state, enabling developers to simulate user interactions and maintain session information.
- Page manipulation: HtmlUnit allows developers to manipulate the HTML content of web pages, modify elements, submit forms, and simulate user interactions.
- Integration with WebDriver: HtmlUnit can be used as a WebDriver implementation, allowing it to work with testing frameworks such as Selenium for web application testing.

3. Natural Language Process

- NLTK (Natural Language Toolkit): NLTK is a Python library that provides a comprehensive suite of NLP tools and resources for tasks like tokenization, stemming, part-of-speech tagging, parsing, and more.
- SpaCy: SpaCy is a Python library known for its efficient and industrial-strength NLP capabilities. It offers fast tokenization, named entity recognition, part-of-speech tagging, dependency parsing, and other NLP features.
- Stanford CoreNLP: Stanford CoreNLP is a suite of NLP tools developed by Stanford University. It includes various components for tasks like tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and more.
- Gensim: Gensim is a Python library focused on topic modeling, document similarity analysis, and natural language processing. It provides implementations of popular algorithms such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

- Apache OpenNLP: Apache OpenNLP is an open-source Java library that offers a wide range of NLP tools and models. It provides support for tasks like tokenization, sentence segmentation, part-of-speech tagging, chunking, named entity recognition, and more.
- Stanford NER (Named Entity Recognizer): Stanford NER is a popular tool for named entity recognition, developed by Stanford University. It identifies and classifies named entities in text, such as names of people, organizations, locations, and more.
- IBM Watson NLU (Natural Language Understanding): IBM Watson NLU is a cloud-based NLP service that provides a range of functionalities, including sentiment analysis, entity recognition, keyword extraction, document classification, and more.
- Hugging Face Transformers: Hugging Face Transformers is a popular library that offers pre-trained models for various NLP tasks, such as text classification, named entity recognition, question answering, machine translation, and more.

4. Data Visualization for Knowledge Graphs

- NetworkX: NetworkX is a Python library primarily used for the creation, analysis, and visualization of complex networks and graphs. It provides basic graph visualization capabilities for knowledge graphs.
- Cytoscape: Cytoscape is an open-source platform for visualizing and analyzing networks. It offers a powerful and user-friendly interface for creating interactive and visually appealing visualizations of knowledge graphs.
- Gephi: Gephi is an open-source software for visualizing and exploring networks. It provides advanced features for creating dynamic and interactive visualizations of knowledge graphs, including options for layout algorithms, filtering, and customization.

- D3.js: D3.js is a JavaScript library for creating data-driven visualizations on the web. It offers a wide range of features and flexibility for creating custom visualizations, including knowledge graphs.
- Graphviz: Graphviz is an open-source graph visualization software. It provides a set of tools and libraries for creating visual representations of graphs, including knowledge graphs. Graphviz uses a graph description language to specify the structure and attributes of the graph.
- Vis.js: Vis.js is a JavaScript library for visualizing interactive networks and graphs. It provides easy-to-use functions and options for creating dynamic and interactive visualizations of knowledge graphs in web applications.